Can Machine Learning Target Health Care Fraud? Evidence from Medicare Hospitalizations

Shubhranshu Shekhar*
Brandeis University
and
Jetson Leder-Luis
Boston University and NBER
and
Leman Akoglu
Carnegie Mellon University

April 9, 2024

Abstract

The US spends more than $4 trillion per year on health care, largely conducted by private providers and reimbursed by insurers. A major concern in this system is overbilling and fraud by hospitals, who face incentives to misreport their claims to receive higher payments. In this work, we develop novel machine learning tools to identify hospitals that overbill insurers, which can be used to guide investigations and auditing of suspicious hospitals for both public and private health insurance systems. Using large-scale claims data from Medicare, the US federal health insurance program for the elderly and disabled, we identify patterns consistent with fraud among inpatient hospitalizations. Our proposed approach for fraud detection is fully unsupervised, not relying on any labeled training data, and is explainable to end users, providing interpretations for which diagnosis, procedure, and billing codes lead to hospitals being labeled suspicious. Using newly collected data from the Department of Justice on hospitals facing anti-fraud lawsuits, and case studies of suspicious hospitals, we validate our approach and findings. Our method provides an 8-fold lift over random targeting of hospitals. We also perform a post-analysis to understand which hospital characteristics, not used for detection, are associated with suspiciousness.

Keywords: Health care, fraud and abuse, machine learning, anomaly detection, explainable AI, Medicare
JEL Codes: I13, C19, D73, K42, M42

*Research reported in this publication was supported by the National Institute on Aging of the National Institutes of Health under Award Number P30AG012810. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We thank the NBER for data access and support and Lowell Taylor for his contributions to earlier stages of the project. We benefited greatly from comments from Anna Zink and participants of the 2023 ASHEcon conference.

1 Introduction

Fraud in health care is hard to detect. Insurers face information asymmetries, where providers know more about the health care delivered than the insurer responsible for paying for that care. Health care providers such as doctors and hospitals face incentives to maximize their reimbursements from health insurance companies, and insurers must largely rely on documentation from providers themselves. This asymmetric information leads to circumstances where unscrupulous providers can choose to commit fraud by manipulating the provided documentation.

These issues are compounded in the US federal health care programs, where the government is the insurer. The US federal government spends over a trillion dollars per year on health insurance, largely paid to private firms, and fraud detection is challenging due to the sheer volume of claims being processed. In 2019, Medicare (the largest of these programs) spent $800 Billion, and even small shares of fraud lead to large losses, taking away funds for valuable care. The US Government Accountability Office (GAO) estimates Medicare improper payments, a measure of mistaken or inappropriately documented spending, in 2019 at $46.2 Billion (U.S. Government Accountability Office, 2020). This problem has gained the attention of Medicare administrators faced with the challenge of detecting and deterring fraud and abuse to ensure the program stays financially solvent (U.S. Department of Health and Human Services, 2022).

Machine learning poses a potential solution to the problem of health care fraud detection, but has been hampered by the challenges of health care claims data, which are highly complex and multi-dimensional. The government observes health care claims including diagnostic, procedural, and billing information, amounting to tens of thousands of potential categorical codes that can be used. Moreover, methods that are based on matching known patterns of fraud, which could be used for supervised machine learning, are biased by the

fact that enforcement is non-random, and will likely fail against the ever-changing nature of fraud in health care.

In this work, we develop and validate new machine learning tools to detect health care overbilling and fraud, which can be used to guide anti-fraud investigations. First, we construct a data formalism for understanding health care claims at the diagnostic-procedural, billing, and spending level, which allows for the detection of rare patterns between types of claims that can be expected to be similar. We then apply recent anomaly detection tools from the computer science literature to detect anomalous providers based on their coding patterns and its effects on hospital spending. These methods rely on the fundamental idea that providers mostly behave like their peers in the absence of fraud, and that deviations from this pattern are more suspicious when they earn the hospital more money. Our approach is unsupervised, i.e. does not need any *a priori* labeled training data, meaning it is not biased by labeled data from past enforcement, going beyond known patterns of fraud.

We apply our method using millions of claims from inpatient hospitalizations, the largest category of Medicare spending, which costs the US government more than $100 Billion per year. Our detection method does not flag any particular hospitalization claim as suspicious, but rather detects hospital-level patterns of care that appear anomalous when considering patient characteristics, medical history, and patterns of behavior by other hospitals when treating similar conditions and patients. This method ranks hospitals in order of their suspiciousness, and is an "explainable" rather than a black-box method, in that it can provide explanations for which types of codes are most anomalous (i.e. potentially misused) for each hospital in our ranking. By ranking hospitals and providing explanations for each ranking, our method speaks to the government's prioritization problem of choosing hospitals for additional scrutiny (such as auditing) with limited enforcement capacity.

We validate our approach with newly collected ground-truth data from the Department

of Justice (DOJ). Using a corpus of thousands of DOJ press releases about fraud, we tag hospitals ever named by the DOJ, and compare these data with our ranking. While only 1 in 20 hospitals nationwide have ever been named in the DOJ Press releases, our ranking substantially improves detection over random sampling: the top 50 hospitals identified by our method contain 21 hospitals named in the same DOJ corpus, an 8-fold lift in detection rate. We note that hospitals highly ranked by our method but not listed by the DOJ are not necessarily false positives; DOJ enforcement depends on opportunity and capacity constraints, providing only partial ground-truth. The DOJ validation resembles positive-unlabeled data (Bekker and Davis, 2020), and the overlap with our method is therefore a lower-bound for detected fraud.

Our algorithm is an ensemble method, utilizing three novel unsupervised detection algorithms that uncover aberrant patterns in care across different levels of claims data, from most fine-grained to the most broad. The first component focuses on the coding behavior within claims, uncovering unusual ICD-10 procedure and diagnosis coding patterns employed by hospitals, which is indicative of manipulation of a patient's codes to garner higher reimbursements. The second component is peer based, focusing on identifying aberrant hospital stay-level billing code (DRG) patterns, compared to peer hospitals that share similar patient populations and distributions of types of care. The third component of the ensemble focuses on hospitals with large observed expenditures conditioned on patient characteristics and medical history, using a regression-based method. To assemble the evidence from these three detection methods together to rank hospitals based on suspiciousness, we utilize instant-runoff voting (Franceschini et al., 2022), which combines information from our different detectors to reach an aggregate ranking. This method follows an iterative procedure to rank the hospital that is most suspicious based on the "vote" across different detectors in each round.

The results of our analysis also provide evidence that is qualitatively consistent with

detecting fraud, rather than singling out legitimate anomalies such as rare or specialty care. We compute the top ICD diagnosis and procedure codes that contribute to identifying hospitals as suspicious. These codes tend not to be rare conditions that are expensive to treat, but rather diagnoses with high payment rates and high ambiguity, indicating they can be more easily manipulated. We also provide case studies of two hospitals to show how our method can be used to dive deeper into data, and show the exact diagnosis and billing codes that make those hospitals suspicious. Finally, we perform an explanatory analysis of the types of characteristics – not used for detection – that are correlated with hospital suspiciousness.

This approach has many potential applications for healthcare policymakers, auditors, and enforcers. While our explanations cannot provide legal-standard evidence of fraud by hospitals, they can be used as starting points that guide deeper investigation such as audits or claim reviews, prioritizing the most suspicious hospitals. Our method can also be readily adapted to detect overbilling in other areas of potentially fraudulent care besides hospitals, such as outpatient claims. Moreover, while the data set on which we build our method is from Medicare, we anticipate our methods will prove valuable to private insurers as well, who face nearly identical challenges in eliminating fraud from private health insurance systems.

This paper proceeds as follows. We describe background and institutions in Section 2 and our data in Section 3. Section 4 provides an overview of our detection methodology, with details in Sections 5, 6, and 7. Section 8 reports the ensemble model detection results, and Section 9 provides a post-analysis that characterizes hospitals with high estimated suspiciousness.

2 Background

In this section, we discuss the institutional details of Medicare fraud. While many of the institutional details about Medicare claims and enforcement are specific to the federal system, the general nature of health care billing is consistent across both publicly funded and private-payer systems.

2.1 Hospital Billing

Medicare implements a Prospective Payment System (PPS) that uses Diagnosis-Related Groups (DRGs) to determine fixed payments for hospital services based on patients' diagnoses, encouraging cost-effective care while ensuring adequate reimbursement. Patients are coded with diagnoses and procedure codes based on the International Classification of Diseases (ICD) system, and then based on this coding, each inpatient stay is classified into one DRG. Because the patient's ICD coding dictates their DRG and ultimately the hospital reimbursement amount, hospital coding decisions directly affect hospital profits. Hospitals do not receive additional reimbursement for providing higher quality services, or a higher volume of procedures for a given patient, although hospital payments are adjusted for high-level factors such as local wage variation and share of medical students trained. MedPAC (2023) presents more details about the hospital prospecitve payment system.

2.2 Health Care Fraud

Fraud in inpatient hospitalization takes many forms. One well-studied form is upcoding, where hospitals miscode patients to higher severity levels of care in order to receive higher reimbursement (Dafny, 2005; Silverman and Skinner, 2004; Becker et al., 2005). A second common issue is lack of medical necessity, where a patient's health conditions do not qualify them for that care (Howard, 2020). Moreover, there is a variety of conduct that can also

qualify as health care fraud, such as providing compensation to providers for referring patients, which qualifies as a kickback.

In this paper, we are largely agnostic to which type of fraud hospitals commit, and instead focus on payment levels. In general, fraud is of greatest concern when it results in higher levels of spending. Our method detects hospitals whose anomalous conduct results in higher payments, which is valuable for detecting hospitals where additional auditing is of highest marginal value.

2.3 Health Care Anti-Fraud Enforcement

The US government undertakes a number of initiatives to detect and deter fraud, waste, and abuse in federally-funded health care spending. Our method, which relies solely on claims data, is complementary to existing methodologies. Private insurers face similar challenges and also work to detect, investigate and enforce against fraudulent providers, although they lack the full weight of the federal investigatory system.

Federal law prohibits Medicare fraud and provides avenues by which fraud can be addressed through criminal and civil enforcement. The federal health care fraud statute provides criminal penalties for those who commit health care fraud, and this enforcement is compounded by criminal enforcement under the anti-kickback statute, as well as the wire fraud and racketeering statutes. Criminal Medicare fraud is prosecuted by the Department of Justice. For a deeper treatment of criminal Medicare fraud, see Eliason et al. (2021).

Civil enforcement for Medicare fraud operates through the False Claims Act, which provides an avenue for whistleblowers to come forward with information about fraud and receive compensation. Whistleblowers file their own cases in federal civil court, and the DOJ has an option to support these cases. Leder-Luis (2023) and Howard (2020) provide more information about the False Claims Act and show that these whistleblowers provide high deterrence effects.

In addition to litigation, administrators use a variety of policy tools to limit health care fraud, waste and abuse. The Office of the Inspector General of Health and Human Services undertakes administrative actions against firms that overbill Medicare. Medicare also has a variety of auditing programs that seek to detect unnecessary or unjustified spending; see Shi (2022) for a description of the Recovery Audit Contractors program. Finally, Medicare uses regulations to target unnecessary spending, such as prior authorization requirements. Some of these regulations combat fraud while others combat waste; see Brot-Goldberg et al. (2022) and Eliason et al. (2021) for a discussion of these regulations.

In addition to the enforcement actions listed above, Medicare and private insurers undertake some data-driven investigatory work in order to detect fraud. These efforts have received little attention in academic work. Medicare claims processors work with contractors called Unified Program Integrity Coordinators (UPICs) (Noridian Healthcare Solutions, 2022) to audit and detect aberrant payments. In addition, Medicare uses a private-public partnership model through the Healthcare Fraud Prevention Partnership to share data between the federal government and private insurers to detect health care fraud with patterns similar across a variety of types of care and different health insurance programs (Healthcare Fraud Prevention Partnership, 2022). When fraud is identified through these data-driven efforts, investigators can refer those cases to the DOJ for civil or criminal prosecution.

In this paper, we curate a list of hospitals that have been subject to DOJ actions at both the criminal and civil level, used for quantitative evaluation of our method. While there are many ways in which hospitals could have been investigated or sanctioned, being named in a DOJ press release validates that the hospital was likely committing behavior that rose to the level of criminal or civil fraud, which represents a true positive. A disclaimer, on the other hand, is that the hospitals subjected to DOJ actions likely constitute only a partial list of all fraudulent hospitals, as other unknown fraud may have gone undetected, which represents a false negative.

2.4 Related Methodological Work

In addition to the economic studies listed above that discuss health care fraud, several data-centric approaches have been explored in the context of Medicare fraud. We refer the reader to Bauder et al. (2017); Kumaraswamy et al. (2022); Joudaki et al. (2015) for detailed surveys on different methods.

In early work, Rosenberg et al. (2000) study upcoding within the claims data. They estimate the probability that a claim has incorrect DRG code, which they further use to identify claims to investigate and audit. Brunt (2011) studies upcoding in the physician office visits data, where he estimates the likelihood of a disease code selected for an office visit to understand upcoding practices. Fang and Gong (2017) find evidence of provider overbilling using inappropriately high number of hours worked to identify outliers.

Recently, Chandola et al. (2013); Suresh et al. (2014) introduce methods for provider profile comparison to spot possible misuses or fraud. These works focus on introducing methods and features to represent hospital profiles for comparison, however, do not present any conclusive results. On the other hand, Bauder and Khoshgoftaar (2018a,b); Herland et al. (2018); Bauder and Khoshgoftaar (2017) utilize publicly available excluded providers to learn models for detection of fraudulent providers. However these approaches rely on the availability of human labeled information on fraudulent information, which is often incomplete and hard to obtain for massive Medicare data.

In contrast to earlier methods, unsupervised and explainable methods for the problem, which are more practical in the real world, have received limited attention. Luo and Gallagher (2010) analyze DRG distributions of hospitals providing services for hip replacements and heart attacks to find upcoding, with the underlying assumption that most hospitals will have similar distributions. Recently, Ekin et al. (2019) learn joint distribution of medical procedures and providers using outpatient data. The joint distribution is used to identify

provider anomalies based on procedure code and usage frequency by the provider.

Most existing research uses only a fraction of the massive Medicare data, relies on labeled data on known fraud, and often does not incorporate an explanation of results that could be useful to guide deeper investigation. Our method builds upon these studies to provide a precise and explainable detection method that does not rely upon the existence of labeled data.

3 Data Description

This study combines data from a variety of sources to detect anomalous hospital spending behavior in Medicare and compare it to ground-truth labeling of hospitals that have faced anti-fraud enforcement.

Our hospital anomaly detection method uses a large-scale dataset of Medicare claims. Data were accessed through a data use agreement with the Centers for Medicare and Medicaid Services, facilitated by the Research Data Assistant Center (ResDAC) and the National Bureau of Economic Research (NBER). These hundreds of millions of observations contain extensive data about each hospitalization and patient in the Medicare system, providing an ideal corpus with which to study hospital behavior.

We consider all patients hospitalized in 2017, and we use data from 2012 through 2016 to construct the patients' medical history. For these years, we use 100% samples of Fee-For-Service institutional Medicare data, including inpatient and outpatient claims, and beneficiary[1] information including demographic information and chronic condition indicators. To further understand a beneficiary's medicare history, we use 20% of samples of

[1]Patient refers to a person receiving health care; beneficiary refers to a person covered by health insurance. Here, they are used interchangeably, as all of our data come from patients who are Medicare beneficiaries.

Table 1: Inpatient data statistics from year 2017

Spending	
Medicare total expenditure(Annual Report, 2022)	$710 billion
Medicare inpatient expenditure	$131 billion
Beneficiaries	
Number of inpatient beneficiaries	6.6 million
Number of inpatient claims	11.2 million
Hospitals	
Number of Hospitals	7,661

carrier files, which describe physician office visits.[2]

Table 1 describes the sample of inpatient hospitalization claims from 2017. We observe 11.2 million claims from 6.6 million beneficiaries representing 7,661 different hospitals. Medicare spent in total $131 billion on inpatient care in 2017, out of $710 billion total reported Medicare spending.

Table 2 describes our sample used to construct patient medical history from 2012 through 2016. We observe nearly a hundred million physician office visits and another hundred million outpatient visits per year, as well as millions of inpatient visits per year. Appendix A provides additional details about the cleaning and use of the Medicare data.

To understand hospital characteristics, we use the Medicare Provider-of-Service files, which contain details on providers such as certification number, name, the type of Medicare services that it provides, and type of ownership (private or public). We can identify patients across files using their unique beneficiary identifiers, and we identify hospitals by their CMS Certification Number (CCN). Further, we separately identify Academic Medical Centers based on their membership to Council of Teaching Hospitals (AAMC, 2022).

[2]20% samples are the largest available for physician office visits.

Table 2: Scale of data from years 2012 to 2016 used to build medical history of patients who are 70 years or older who appear in the inpatient claims from year 2017. The number in each cell is in millions.

	2012	2013	2014	2015	2016
Physician visits	94.7	100.2	102.8	107.7	114.2
Outpatient visits	81.5	87.3	90.9	96.8	104.3
Inpatient visits	4.0	4.2	4.4	5.1	5.8

The federal Department of Justice (DOJ) publishes press releases when fraud is identified in civil or criminal lawsuits, in order to inform the press and the public as well as deter future fraudulent behavior. To evaluate our automated detection of suspicious hospitals, we utilize these press releases related to Medicare from the DOJ. To that end, we scraped from the DOJ website thousands of press releases that contain the word 'Medicare'. Each press release corresponds to a case that the Department of Justice was involved with, often at the time of settlement. Using partial name matching, we tag the hospitals that appear in this corpus. This also accounts for hospital chains, when the chain name appears in both the hospital name and the press release.

As the DOJ lacks both the capacity and the information to prosecute all Medicare fraud, the press releases provide only a partial list of hospitals that have engaged in fraudulent behavior. We can consider this a form of positive-unlabeled data: while we can identify firms that have been named in a press release as having likely committed fraud, firms that are *not* named are not necessarily above suspicion. In general, we consider the DOJ data as a partial ground truth, as a sensible though possibly incomplete way to measure whether the firms identified by our metric are validated as the firms that have at some point committed fraud, and therefore deserve additional scrutiny. Appendix B provides

additional details about the collection and cleaning of the DOJ corpus.

4 Method Overview

Medicare claims data contain many levels of detailed information about hospital claims, including procedure and diagnosis codes, claim-level billing codes, and patient characteristics, which provide opportunities for modeling the fraud detection problem in various ways. For example, a hospital can be represented by the frequency of ICD (diagnosis and procedure) codes used in its claims, the DRG (billing) codes associated with its claims, or by the characteristics of the patient populations that it serves. Each data modality presents us with a specific perspective of the data, which when combined allow us to learn comprehensive hospital behavior which reveal information that cannot be completely uncovered based on only one aspect of the data.

In this work, our goal is to estimate a suspiciousness ranking for hospitals. We use an unsupervised *multi-view* anomaly detection approach, suitable for the underlying multi-modal data. Each view (or base detector) presents itself as a different model of the anomalies, operating on a different data representation. As such, each can be seen as providing evidence that corresponds to a particular reason for detection. The explanation provided by each detector provides a unique perspective into suspicious behavior. Collectively, the evidence from these base detectors can be assembled systematically into an ensemble detection method.

Ensemble methods utilize multiple base detectors, where under certain accuracy and diversity conditions, they are to obtain better performance than the constituent base detector alone and produce more robust results (Aggarwal and Sathe, 2017). Diversity is an important property of ensemble methods, which ensures that the base detectors make independent errors that cancel out when aggregated. Therefore, various approaches have been

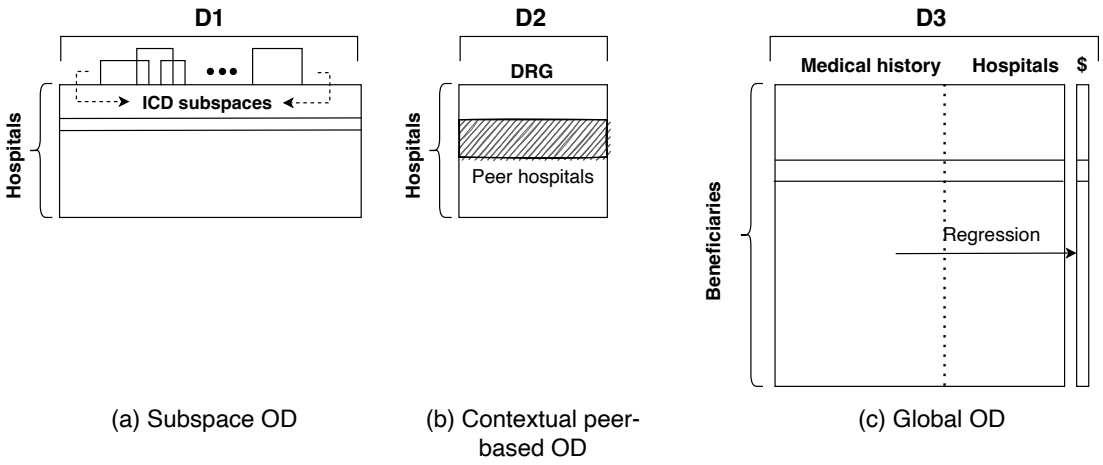


Figure 1: Multi-view anomaly detection on different Medicare data modalities – D1, D2, and D3. Model (a): Local (in ICD codes) detector in the very high dimensional ICD code frequency representation of hospitals. It explains anomalies based on feature importance, i.e. with respect to specific ICD codes. (b): Local and contextual (peer-based) detector based on comparing DRG frequency distributions. It provides a contrastive explanation in terms of excess cost of treatment when compared to peers. Model (c): Global detector based on fixed effects regression model. The coefficient of a hospital is an indicator of excess cost of care at the hospital.

proposed toward promoting ensemble diversity (Kuncheva and Whitaker, 2003; Nam et al., 2021). In essence, our approach utilizes the diversity of the underlying data representations to induce diversity in the ensemble.

Figure 1 shows the different Medicare data modalities we consider and provide a high level description of the corresponding base outlier detection (OD) model that utilizes it. The first OD model, Figure 1 (a), performs outlier detection among hospitals as represented by the frequency of ICD codes used in their claims (denoted **D1**). Anomalous coding may be associated with only a few ICD codes (i.e. features) at a time, rather than all. Therefore, this model is a feature subspace detector, finding outliers locally in subsets of features. The second OD model, Figure 1 (b), performs contextual detection, identifying hospitals that

behave differently from their peers. Behavior is captured by the frequency distribution of the DRG codes assigned to each hospital's claims (denoted **D2**). Here, we recognize the heterogeneity among hospitals and compare a hospital's behavior locally, i.e. in the context of its peers with similar characteristics. Finally, the third OD model, Figure 1 (c), is set up as a global regression onto cost per beneficiary (target variable) from data (denoted **D3** on the figure) reflecting a beneficiary's medical history and the hospitals that they visited.

In addition to detection, our proposed models can provide explanations for their flagged anomalies. This is especially important in the absence of any ground-truth labels in practice, which allows investigators to determine why a hospital is ranked as an outlier, and facilitating decisions such as whether to conduct additional investigation or to audit. By capitalizing on different data representations, our method leads to different explanations with each OD model, enabling a multi-view reasoning. Specifically, in Figure 1, the selected subspace in the first OD model (a) quantifies feature (i.e. ICD code) importance, and can explain each flagged anomalous hospital based on the specific ICD codes that they use differently in their claims. The second OD model (b) provides contrastive explanations, through comparing DRG frequencies of a hospital to those of their peers. As the DRG code of a claim dictates cost, differences in the DRG coding distribution can be directly translated to excess cost of treatment. Importantly, the explanation can pinpoint which DRGs are most contributing to large excess cost of a hospital, facilitating auditing. The regression coefficient associated with a hospital in our last OD model (c) is a direct indicator of excess spending at the hospital.

To arrive at a final anomalous ranking based on different modalities, we combine the rankings from individual detectors to capture the agreement among them. We use instant-runoff voting (Franceschini et al., 2022) to combine information from our different detectors. Our ensemble approach allows us to gather evidence from multiple models, but can also be "unrolled" to provide explanations to each flagged anomaly by each detector in the

ensemble.

The following three sections are organized to present the details of our three detection models, in terms of data set up, detection methodology and explanation. Then, we present the results of our aggregate ranking in Section 8.

5 Detection through ICD Coding Subspace Analysis

ICD codes are used by health care providers to characterize a patient's medical condition and treatment. The US uses ICD Version 10 codes, which were developed by the World Health Organization and can be used to designate the universe of medical issues and procedures. ICD codes encode hospital assessment of a patient based on their reason of visit to the hospital and their medical conditions, and primarily reflect the diagnoses and applied procedures for treatment. For Medicare billing, the assigned ICD codes are then used as input to a "grouper" software used by hospital billers that assigns a diagnostic code (DRG) based on the hospital's findings as indicated by the assigned ICD codes. As discussed above, in the Medicare PPS, the DRG code determines the reimbursement level. Consequently, ICD coding presents opportunities for miscoding, as hospitals may try to achieve a more expensive DRG code to obtain higher reimbursement. Therefore, the objective of our ICD coding based analysis is to understand hospital coding practices that could reveal the coding patterns applied by hospitals engaging in fraudulent behavior.

5.1 Data Setup

Hospital representation.

We use inpatient claims from the year 2017, gathering ICD codes from each claim. We represent hospitals through their reported ICD codes, including diagnostic and procedure codes.

Importantly, since hospitals have a choice of ICD codes, we must also account for ICD *code substitutability*, where a slightly similar ICD code could be used instead to yield higher reimbursements. To capture code substitutability, we estimate the semantic similarity of the description of each code within each chapter of the ICD code hierarchy. Here, the description of each ICD code is constructed by concatenating its text description to the description of its ancestor codes within the ICD hierarchy. Then, pairwise Jaccard distance is computed between the descriptions of the codes and the hospital representation is updated using the ICD code similarity.

For example, the description of ICD code J45.20 under chapter X is constructed by concatenating the descriptions of J00-J99 chapter, J40-J47 block, J45, and then the ICD code J45.20 resulting in the description given as "Diseases of the respiratory system – Chronic lower respiratory diseases – Asthma – Mild intermittent asthma uncomplicated." This code will be similar to other codes that contain the word "asthma" or "respiratory sytstem". This representation also ensures that codes nearby in the ICD hierarchy have somewhat similar text descriptions and are therefore near each other in Jaccard distance.

Formally, let $\boldsymbol{X}^{ICD} \in \mathbb{R}^{N_H \times M_H}$ be the matrix representation of N_H hospitals in terms of M_H-dimensional ICD codes in which the entries depict the total code usage count by the hospital, and $\boldsymbol{J} \in \mathbb{R}^{M_H \times M_H}$ be the ICD substitutability matrix consisting of pairwise Jaccard similarities. Then, the provider representation $\boldsymbol{X}^{ICD_{sim}} \in \mathbb{R}^{N_H \times M_H}$ after incorporating the code substitutability is given as $\boldsymbol{X}^{ICD_{sim}} = \boldsymbol{X}^{ICD} \times \boldsymbol{J}$, which re-distributes each code's frequency to substitutable ICD codes that are not directly reported in the claims data.

We note that $\boldsymbol{X}^{ICD_{sim}}$ is very high dimensional ($> 40,000$ features). However, anomalous coding of a claim is likely covert and associate with only a few ICD codes. Therefore, we employ a feature *subspace* based detector for finding outliers locally among subsets of ICD codes. Figure 1(a) shows this setup.

5.2 Detection Model

We employ a suite of subspace outlier detectors on the high dimensional hospital representation $\boldsymbol{X}^{ICD_{sim}}$ to find hospitals deviating from the majority coding practices within certain ICD subspaces. As we are interested in ICD subspaces that are relevant for a variety of aberrant hospital practices, we utilize an ensemble of subspace detection methods that are effective on high dimensional data. In the same spirit as with our overall approach, the ensemble allows us to examine multiple diverse subspaces as each subspace detection method implements a different methodology for exploring candidate subspaces. In particular, our subspace ensemble uses five different state-of-the-art methods that we describe briefly below.

Subspace outlier detection.

While we represent a hospital in the high dimensional ICD space, the abnormal or aberrant behavior may be reflected only in a small, locally relevant subset of codes – that is, only certain codes will be fraudulently or suspiciously substituted. Each OD algorithm in the ensemble explores local subspaces differently to provide evidence from diverse subsets. To that end, our OD model consists of the following subspace detectors: SOD (Kriegel et al., 2009), IF (Liu et al., 2008), RRCF (Guha et al., 2016), LODA (Pevnỳ, 2016), RSHash (Sathe and Aggarwal, 2016). Details of each method are included in Appendix C.

We apply the above methods to $\boldsymbol{X}^{ICD_{sim}}$, the ICD representation of hospitals, and identify the hospitals that behave abnormally in various subspaces as explored by the algorithms.

Anomaly scoring.

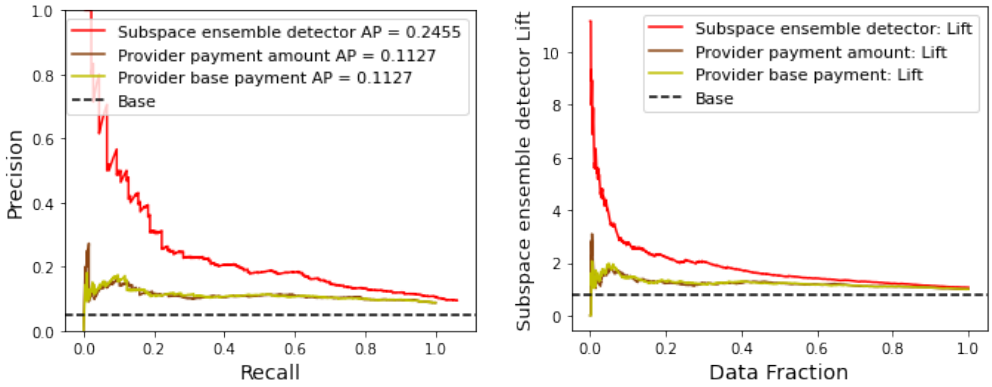
Each subspace algorithm assigns an anomaly score to each hospital. The scores have different scale and semantics (path length, likelihood, etc.), and thus are not directly comparable across the methods. Therefore, we aggregate the *ranking* of hospitals based on individual

scoring of each subspace method. We use the instant-runoff voting technique (details in Section 8) for rank aggregation from different subspace algorithms, and provide the final ranking of hospitals by anomalousness across all subspaces.

5.3 Model Explanation

We explain the ranking of a subspace detector using Shapley Additive Explanation values (SHAP values), introduced in Lundberg and Lee (2017) and Lundberg et al. (2020). SHAP values estimate feature importance by approximating the effect of removing each feature from the model as the average of differences between the predictions of a model trained with and without the respective feature. We regress the anomaly scores from a subspace detector onto the ICD representation of hospitals, and then estimate the SHAP values under the regression model. The feature contributions for each observation find the most important codes that affect the anomaly score significantly. This helps us find ICD codes that are contributors to a hospital being ranked as an outlier.

Further, we provide dollar amount characterization of important features (ICD codes). In practice, combinations of ICD codes are used to determine DRG claim codes, which correspond to payment. For our purposes, each ICD code is mapped to the most frequent DRG code assigned for the given ICD code within the inpatient claims. Since DRG codes are determinants of the payment for care, through this most-frequent DRG mapping, we associate dollar amount of reimbursement to ICD codes. This lends itself to understanding the dollar impact of an important ICD code for an anomalous hospital as explained by SHAP feature importance values.



(a) Precision-recall curve

(b) Lift curve

Figure 2: We report (a) Precision-Recall curve and (b) Lift curve for hospital ranking produced by our ICD-10 subspace outlier detector ensemble against two simple baselines that rank the hospitals based on (1) average total claim amount and (2) average base payment amount. Dashed horizontal line 'Base' depicts the random ranking.

5.4 Evaluation

To evaluate the effectiveness of our ranking, we use the partial list of known fraudulent hospital based on the DOJ press releases described in Section 2.3, and we compare our suspicious hospitals to known fraudulent hospitals. We quantitatively evaluate the targeting of fraudulent hospitals using two ranking quality metrics, namely a Precision-Recall (PR) curve, and a Lift curve.

The PR curve depicts the positive predictive value (precision) on the y-axis versus the true positive rate (recall) on the x-axis. Those responsible for detecting and enforcing fraud in Medicare have a limited budget, and therefore have to select limited targets for deeper investigation; therefore, a high precision at the **top** of the ranked list is most useful. Similarly, the lift curve measures the targeting effectiveness on y-axis when compared to a random baseline as we move along varying fractions of the ranking on the x-axis.

Figure 2 reports the performance of our subspace OD model in terms of the PR and Lift

curves, using the DOJ ground truth. The ICD subspace model ranking alone is at least 2× better at targeting fraudulent hospitals compared to our two baselines, respectively based on total claim payment and base payment amounts. Our method substantially outperforms random auditing or even detection based strictly on payment amounts.

6 Expenditure-Based Detection with Peer Analysis

Our second model is based on peer-based excess spending detection and examines the coding decisions of hospitals as compared to similar "peer" hospitals that treat similar populations. In short, we identify hospitals who are exposed to the same patient population but manage to assign more expensive DRG billing codes.

The objective of the peer-based analysis is uncovering the local patterns of spending behavior among a *related* group of hospitals called peers, and identifying hospitals deviating from the group's expected behavior. We utilize the inpatient claims to create a profile for each hospitals under two complementing data modalities, based on: (1) type of services provided by the hospital, and (2) the patients' chronic condition profiles served by a hospital. We then find groups of related hospitals based on the similarity of this representation.

To identify a locally aberrant behavior, each hospital is represented in terms of its DRG frequency distribution, which determines spending. Then, the DRG representation of a given hospital is compared to the summary DRG distribution of their peers. Figure 1(c) visualizes this setup. The hospitals are then ranked in order of their deviation from group behavior in terms of DRG-based spending.

6.1 Data Setup

Hospital representation.

We construct hospital profiles to capture the nature of services provided (using major diagnostic categories or MDCs), the characteristics of patient population served (using patient's chronic conditions), and encoding practices that drive spending for treatment (using DRGs). The details of the representation are provided in Appendix D.

6.2 Detection Model

Peer identification.

We create peer groups of hospitals that share similarities in the type of services provided or the patient population served.

Let $\boldsymbol{v}_j$ denote the representation for hospital j; either based on the type of services profile using MDC codes or based on the patient population profile using chronic conditions of patients. We note that the hospital representations are frequency distributions, as they depict normalized counts. Therefore, to measure the similarity between two hospitals j and k, we use the Hellinger distance for probability distributions, which is an upper bound on the total variation distance (Bar-Yossef et al., 2004), given as

$$d_{jk} = \frac{1}{\sqrt{2}} \cdot \|\sqrt{\boldsymbol{v}_j} - \sqrt{\boldsymbol{v}_k}\|_2 \tag{1}$$

We examine the distribution of pairwise similarity values to decide on a threshold τ to include only the most similar hospitals in a hospital's peer group.

For each hospital j, the hospitals with similarity to j above τ constitute j's peers, denoted $\mathcal{P}_j$. Notice that the peers are specified for each hospital separately, rather than using any clustering algorithm. This allows us to create compact peer groups of varying sizes.

We note that fixing the peer group size would be a subpar alternative, since j's group may then include distant hospitals as peers, skewing the representative summary statistics of the group that j is compared to.

Anomaly scoring.

In the Medicare PPS, the reimbursement amount for treatment is directly based on the assigned DRG code to a claim. Therefore, for anomaly scoring, we utilize the hospital representations over DRG codes from the inpatient claims, which consist of the normalized counts of the DRG codes used by a provider. In short, this detection mechanism assumes that hospitals who treat similar patient populations, or provide care for similar categories of illnesses and injuries, should have similar DRG distributions.

For each hospital, we create a peer group summary in terms of distribution over DRG codes among all peers, weighted by their similarity to the hospital under consideration. Let $\boldsymbol{v}_j^{DRG}$ be the DRG distribution for hospital j with n_j claims, and $\boldsymbol{q}_j^{DRG}$ be the summary DRG distribution based on hospital j's peers, defined as follows.

$$\boldsymbol{q}_j^{DRG} = \tfrac{1}{Z} \textstyle\sum_{k \in \mathcal{P}_j} n_k \times (1 - d_{jk}) \times \boldsymbol{v}_k^{DRG} \quad \text{where } \mathcal{P}_j = \{\, k \mid (1 - d_{jk}) \geq \tau \,\} \text{ and } Z = \textstyle\sum_{k \in \mathcal{P}_j} n_k \times (1 - d_{jk}) \quad (2)$$

Next we tie the DRG usage frequencies to average dollar amount spending by Medicare, as the former dictates the latter. $Cost(c)$ denotes the average base price of DRG code c computed from the inpatient claims data from the year 2017. Then, the excess spending for treatment per claim on average for provider j is given as follows:

$$ExcessSpending_j \;=\; \sum_{c \in DRGs} Cost(c) \;\times\; (\boldsymbol{v}^{DRG}_{j,\,\text{index}(c)} - \boldsymbol{q}^{DRG}_{j,\,\text{index}(c)}) \quad (3)$$

where $\boldsymbol{v}^{DRG}_{j,\,\text{index}(c)}$ is the frequency corresponding to DRG code c in the DRG representation $\boldsymbol{v}_j^{DRG}$ for provider j, and $\boldsymbol{q}^{DRG}_{j,\,\text{index}(c)}$ denotes that for DRG code c in the peer group summary representation $\boldsymbol{q}_j^{DRG}$. In short, this amount computes how much more a hospital spends

because they use a different set of DRG codes than their peers, based on the average price of those DRGs.

The calculated $ExcessSpending$ amount is the anomaly score based on which the hospital are ranked, as it depicts the average spending discrepancy for a hospital when compared to peers of the given hospital. Since we create two different peer groupings – one based on services provided, and another based on patients served – we obtain two rankings, later combined through instant-runoff voting (Section 8).

6.3 Model Explanation

The peer based OD model's anomaly score is the estimated excess spending, which is directly interpretable as the extra dollar amount a hospital charges on each claim on average as compared to what would be expected from other similar hospitals. Further explanation can be provided for a top-ranked hospital by contrasting their frequency distribution over DRG codes against their peers. This allows auditors to have a contrastive understanding of DRG codes used by similar hospitals, and to pinpoint to specific DRGs with large frequency discrepancies. Direct usage comparison of individual DRGs could point to specific codes that contribute most to the overall spending at a hospital, and guide a deeper investigation of the claims associated with those specific DRG codes.

6.4 Evaluation

We have included the evaluation details in Appendix E. Additionally, through case studies in Appendix G, we report qualitative results and provide peer-based explanations and insights into top flagged providers after aggregating evidences from different OD models.

7 Expenditure-Based Detection with Massive Fixed-Effect Regression

As a third and final detector, we consider a hospital-level analysis of expenditure to understand which hospitals are associated with high spending on a beneficiary's hospitalization. The incentive of providers who commit fraud is to receive higher reimbursement, and so unexplained high expenditure is potentially suspicious. Our design detects high expenditures that are unexplained by a patient's medical history, which could reflect unnecessary or excessive billing. While any individual patient may receive entirely necessary high levels of care – for example, in response to a severe accident – when a hospital's patient population consistently shows expensive, unexplained high expenditure, this may be indicative of fraud or waste.

Our design considers expenditure as a function of a patient's medical history. We collect each beneficiary's medical history, using claims from physicians office visits, hospital outpatient visits, and hospital inpatient visits over a five-year period before the target year. The outcome variable of the regression is the base claim amount per beneficiary per hospital visited in the current year. Below we provide our model specification, and in Appendix F, we give detailed description of data, algorithm and anomaly scoring.

Regression model specification for expenditure.

Given (i) patient representation $\boldsymbol{X} \in \mathbb{R}^{N \times M}$ for N patients, each with a M-dimensional representation of historical medical profile based on the last five years (2012–2016), and (ii) the total base payment Y in year 2017; the specification for expected treatment expenditure prediction is as follows.

$$Y_i \;=\; \beta_0 \;+\; \boldsymbol{X}_i\,\boldsymbol{\beta} \;+\; \sum_j \alpha_j\, H_{j,i} \;+\; \epsilon_i \;\;, \tag{4}$$

where Y_i is the total base payment expense for a patient i in 2017; $\boldsymbol{X}_i$ is the patient representation for i, $\boldsymbol{\beta}$ depict regression coefficients associated with patient medical profiles and locations, $H_{j,i}$ is associated with an inpatient Medicare hospital j which contains total count of visits to j if patient i visited the hospital and 0 otherwise, and α_j's depict the hospital fixed effect regression coefficients.

8 Aggregate Provider Ranking

Each outlier detection model presented above is a component of our ensemble method that considers different aspects of the data and creates a ranked list of providers. This ensemble method is designed to handle multi-view Medicare data, where different features of the data can be used to evaluate different aspects of suspiciousness. The goal of the ensemble is a single suspiciousness ranking for all providers.

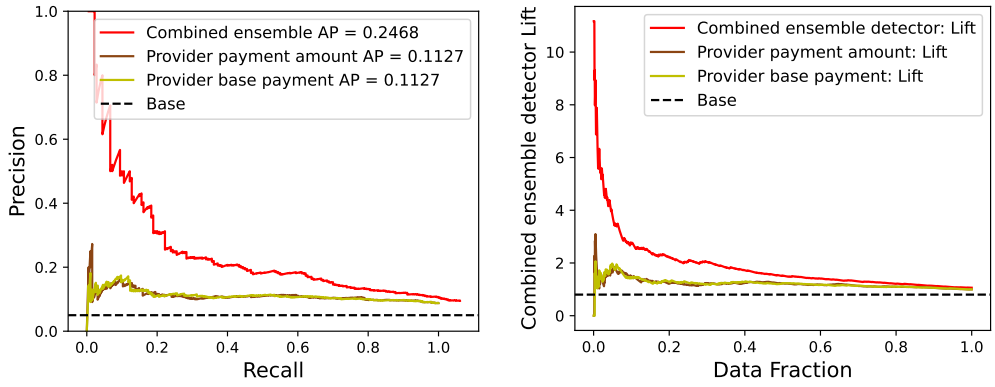
To arrive at the final ranking for auditing, we merge multiple rank lists into a single ranking using instant-runoff voting (IRV). Our goal is to present the aggregate ranking that is most representative of the component models. IRV combines results across rankings in a way that best reflects the information contained across multiple models (Franceschini et al., 2022).

The rank aggregation proceeds in an iterative manner, where each round utilizes the IRV procedure to find a "winner" (in our case, most suspicious hospital). In each round, votes are counted for each component ranking's first choice, and a hospital with a majority of votes is then ranked at top in our aggregate ranking. The rank lists across models are updated to drop the selected hospital in this round, and the IRV procedure is repeated with updated rank lists in the subsequent rounds to arrive at an aggregate ranking.

In our implementation, we aggregate 8 different rankings across our 3 OD models; five from different subspace OD algorithms on the ICD data, two from the peer-based model

utilizing the two separate similarity measures (patient populations and categories of care), and one from the regression model. Next, we show the effectiveness of our final aggregate ranking for identifying fraudulent hospitals in the Medicare system through quantitative evaluations.

Additionally, in Appendix G, we provide qualitative evaluations through case-studies on suspicious providers from our method, highlighting some of the salient aspects for the fraud detection task. We show how our method highlights parts of data from that contributed most to the ranking of a hospital as suspicious, which can assist in the process of auditing or deeper investigation. Appendix H goes further, highlighting the exact ICD codes that greatest contribute to our top-ranked provider suspiciousness. An examination of the codes indicates that providers are not being flagged for treating rare diagnoses, but rather for using diagnosis and procedure codes that reflect *ambiguity* and are relatively highly paid. The intentional use of ambiguous diagnosis codes may reflect an attempt to reclassify patients into more obscure diagnoses to achieve a higher-paid DRG code.



(a) Precision-recall curve

(b) Lift curve

Figure 3: We report the performance of the final ranking of providers as aggregated from 8 rankings based on 3 different OD models. Note that aggregated ranking improves over the ranking by individual constituent experts. The proposed ensemble is on average 4× better than the random targeting of hospitals for auditing.

8.1 Quantitative Evaluation

Figure 3 shows the evaluation of our aggregate ranking of hospitals using a PR curve and a Lift curve. The aggregate ranking is compared to intuitive baselines that rank hospitals based on their average reimbursements, or random auditing. Our aggregate ranking is able to target fraudulent hospitals on average twice as better when compared to the baseline ranking—note the area-under-curve, or average precision (AP) values on legend Figure 3(a).

While only 1 in 20 hospitals are named in the DOJ Press releases, the top 50 hospitals identified by our aggregate ranking contain 21 hospitals named in the DOJ corpus. That is an 8-fold lift in detection rate considering the evaluation at top 50 hospitals, with an average of 4-fold lift over random/by-chance targeting across varying data fractions as seen in Figure 3(b). Importantly, our ground-truth consists only of hospitals named in the DOJ corpus, while there may be others with yet unidentified fraudulent practices – and therefore, our list can be used to find other hospitals not yet identified as fraudulent.

For robustness, we repeat our method on (i) claims originating from emergency room (ER), and (ii) claims excluding Medicare Advantage patients. Appendix J presents the method and results for ER data. This is a different set of patients and claims, who are less likely to have selected the hospital themselves, and therefore this addresses potential selection bias, wherein sicker patients choose more sophisticated hospitals. Despite the sample limitation, we still achieve a strong improvement over random targeting, 2.5× lift among the top 50 hospitals. Appendix L presents the method and results after excluding patients with any Medicare Advantage (HMO) coverage. These patients are covered by third parties and therefore we may not observe their full medical history. Among this sample, our method achieves 5× lift over random targeting among the top 50 hospitals.

Statistical Significance of Ranked Results

We use the $\sqrt{\epsilon}$ test proposed by Chikina (Chikina et al., 2017) to evaluate the statistical significance of the ranked results. The test is based on a rigorous Markov chain framework, namely comparing our ranked result with a random ranking from a well-defined, valid claims data Markov chain. A state in our Markov chain represents possible claims data resulting from a beneficiary seeking potential treatment at a nearby hospital. We provide the details of the test and validity constraints in Appendix K. Note that due to the scale of our data, running the validity checks while constructing the Markov chain is non-trivial, details in Appendix K. We run the test on the Markov chain (with $k = 2^{30}$ steps) obtained from our experiment data. Our ranked results are statistically significant at the 5% level, $p < 0.05$.

9 Characterizing Outlier Providers

In this section, we examine the covariates of hospitals to understand the factors that characterize an outlier hospital as detected by our model. The covariates used in the analysis depict various hospital characteristics such as hospital rating, number of unique patients served, ownership type, location, and length of stay for an inpatient visit. These features are derived from publicly available information for all Medicare hospitals and importantly are *not* included in the data used for detection.

Understanding the factors that drive outlier hospital behavior is crucial for improving the health care sector. Extensive policy reforms seek to shape the structure of the health care market, increasing regulations on providers deemed to be harmful or inefficient. By characterizing the nature of hospitals deemed suspicious by our metrics, we hope to contribute to the ongoing literature that evaluates how various interventions – for example, those targeting for-profit care – can affect fraudulent behavior.

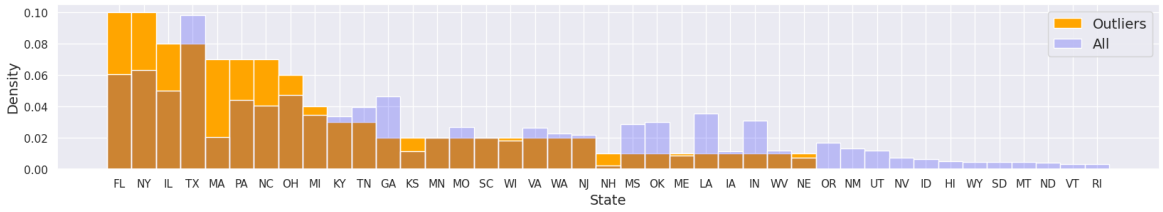


Figure 4: Comparison of Distributions over States for Outliers and All Hospitals

Figure 4 compares the distributions over states where a hospital is located. Outlier hospitals are more likely to be from states Florida, New York, Illinois, and Massachusetts, and less likely to be from Texas and Georgia. This is also corroborated by the DOJ cases, where about 15% of the named hospitals are based in Florida.

(a) Average Length of Stay

(b) #Unique Patients Served

Figure 5: Comparison of Distributions over Covariates for Outliers and All Hospitals

Figure 5 compares the distributions across average length of stay and number of unique patients served. Ranked outlier hospitals keep inpatients longer as compared to other hospitals. This could be to justify the usage of costlier DRGs, or driven by ranked outlier hospitals receiving sicker patients; however, our metrics control for patient health characteristics. Additionally, top ranked fraudulent hospitals serve more unique patients on average. Since a large fraction of our top ranked hospitals are also named by the DOJ, it may indicate that a greater number of unique patients may provide more opportunity for perturbations in diagnosis coding resulting in higher reimbursements, or it could reflect the fact that our outliers are largely urban hospitals. We also include results for categorical

covariates e.g. hospital rating, ownership type, location type in Appendix I.

10 Conclusion

The unsupervised ensemble method introduced in this work provides a new data-driven approach to identifying health care fraud using massive claims data. Our approach uses different data modalities – including patient medical history, provider coding patterns, and provider spending – to detect anomalous behavior consistent with fraud and abuse. Besides detection, the methodology offers interpretability, model-specific explanations pinpoint specific ICD and DRG codes associated with excess spending at a provider. Finally, our method allows us to characterize the types of providers most likely to be ranked as suspicious, which may be useful for guiding anti-fraud policy more broadly.

Our method substantially outperforms baseline algorithms. We combine evidence from multiple unsupervised outlier detection algorithms that use different types of global and local analysis to create a final ranking of suspiciousness, based on Medicare data. While only 1 in 20 hospitals are named in our DOJ ground truth data as fraudulent, 21 of our top ranked 50 hospitals are in the same corpus, achieving an 8-fold improvement in detection rate.

Medicare spends over a hundred billion dollars per year on hospitalizations, and the federal government has limited enforcement capacity. We believe our findings are *per se* interesting, because they help pinpoint fraud by private firms against the government in a way that could be used to improve public spending.

Our method has natural extensions beyond Medicare and beyond hospitalizations. We believe that the same method will prove useful in detecting fraud against private insurers, who face many of the same issues. Private insurers spend hundreds of billions of dollars per year on reimbursing care, and even small shares of fraud can be very expensive. Our

detection algorithm can be used to guide auditing by identifying which providers are committing the most egregious behavior. Our method also has a natural extension to Medicaid, the federal-state partnered low-income subsidy program, which spends an additional $400 Billion per year on health care. With health care spending at 19.7% of US GDP Centers for Medicare & Medicaid Services (2022), tools for detecting health care fraud can find wide-ranging use.

References

AAMC. Council of teaching hospitals and health systems (coth), 2022. URL `https://www.aamc.org/career-development/affinity-groups/coth`.

C. C. Aggarwal and S. Sathe. Outlier ensembles: An introduction. *Springer*, 2017.

Annual Report. 2022 annual report of the boards of trustees of the federal hospital insurance and federal supplementary medical insurance trust funds., 2022. URL `https://www.cms.gov/files/document/2022-medicare-trustees-report.pdf`.

Z. Bar-Yossef, T. S. Jayram, R. Kumar, and D. Sivakumar. An information statistics approach to data stream and communication complexity. *J. Computer and System Sciences*, 68(4):702–732, June 2004. doi: 10.1016/j.jcss.2003.11.006. (Preliminary Version in *43rd FOCS*, 2002).

R. Bauder and T. Khoshgoftaar. Medicare fraud detection using random forest with class imbalanced big data. In *2018 IEEE international conference on information reuse and integration (IRI)*, pages 80–87. IEEE, 2018a.

R. Bauder, T. M. Khoshgoftaar, and N. Seliya. A survey on the state of healthcare upcoding fraud analysis and detection. *Health Services and Outcomes Research Methodology*, 17(1):31–55, 2017.

R. A. Bauder and T. M. Khoshgoftaar. Medicare fraud detection using machine learning methods.

In *2017 16th IEEE international conference on machine learning and applications (ICMLA)*, pages 858–865. IEEE, 2017.

R. A. Bauder and T. M. Khoshgoftaar. The detection of medicare fraud using machine learning methods with excluded provider labels. In *The Thirty-First International Flairs Conference*, 2018b.

D. Becker, D. Kessler, and M. McClellan. Detecting medicare abuse. *Journal of Health Economics*, 24(1):189–210, 2005.

J. Bekker and J. Davis. Learning from positive and unlabeled data: A survey. *Machine Learning*, 109(4):719–760, 2020.

Z. Brot-Goldberg, S. Burn, T. Layton, and B. Vabson. Rationing medicine through bureaucracy: authorization restrictions in medicare. Technical report, Working Paper, 2022.

C. S. Brunt. Cpt fee differentials and visit upcoding under medicare part b. *Health economics*, 20(7):831–841, 2011.

Centers for Medicare & Medicaid Services. Nhe fact sheet, 2022. URL `https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NHE-Fact-Sheet`.

V. Chandola, S. R. Sukumar, and J. C. Schryver. Knowledge discovery from massive healthcare claims data. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1312–1320, 2013.

M. Chikina, A. Frieze, and W. Pegden. Assessing significance in a markov chain without mixing. *Proceedings of the National Academy of Sciences*, 114(11):2860–2864, 2017.

L. S. Dafny. How do hospitals respond to price changes? *American Economic Review*, 95(5):1525–1547, 2005. URL `https://www.aeaweb.org/articles?id=10.1257/000282805775014236`.

DOJ Settlement. 32 hospitals to pay u.s. more than $28 million to resolve false claims act allegations related to kyphoplasty billing, 2015. URL https://www.justice.gov//opa/pr/32-hospitals-pay-us-more-28-million-resolve-false-claims-act-allegations-related-kyphoplasty.

DOJ Settlement. Northern ohio health system agrees to pay over $21 million to resolve false claims act allegations for improper payments to referring physicians, 2021. URL https://www.justice.gov//opa/pr/northern-ohio-health-system-agrees-pay-over-21-million-resolve-false-claims-act-allegations.

T. Ekin, G. Lakomski, and R. M. Musal. An unsupervised bayesian hierarchical method for medical fraud assessment. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 12(2):116–124, 2019.

P. J. Eliason, R. J. League, J. Leder-Luis, R. C. McDevitt, and J. W. Roberts. Ambulance taxis: The impact of regulation and litigation on health care fraud. Technical report, National Bureau of Economic Research, 2021.

H. Fang and Q. Gong. Detecting potential overbilling in medicare reimbursement via hours worked. *American Economic Review*, 107(2):562–591, 2017.

F. Franceschini, D. A. Maisano, and L. Mastrogiacomo. Ranking aggregation techniques. In *Rankings and Decisions in Engineering*, pages 85–160. Springer, 2022.

S. Guha, N. Mishra, G. Roy, and O. Schrijvers. Robust random cut forest based anomaly detection on streams. In *International conference on machine learning*, pages 2712–2721. PMLR, 2016.

A. Gupta, S. T. Howell, C. Yannelis, and A. Gupta. Does private equity investment in healthcare benefit patients? evidence from nursing homes. Working Paper 28474, National Bureau of Economic Research, February 2021. URL http://www.nber.org/papers/w28474.

Healthcare Fraud Prevention Partnership. Healthcare fraud prevention partnership, 2022. URL https://www.cms.gov/hfpp.

M. Herland, T. M. Khoshgoftaar, and R. A. Bauder. Big data fraud detection using multiple medicare data sources. *Journal of Big Data*, 5(1):1–21, 2018.

D. Howard. False claims act liability for overtreatment. *Journal of Health Politics, Policy and Law*, 45(3):419–437, 2020.

H. Joudaki, A. Rashidian, B. Minaei-Bidgoli, M. Mahmoodi, B. Geraili, M. Nasiri, and M. Arab. Using data mining to detect health care fraud and abuse: a review of literature. *Global journal of health science*, 7(1):194, 2015.

H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek. Outlier detection in axis-parallel subspaces of high dimensional data. In *Pacific-asia conference on knowledge discovery and data mining*, pages 831–838. Springer, 2009.

N. Kumaraswamy, M. K. Markey, T. Ekin, J. C. Barner, and K. Rascati. Healthcare fraud data mining methods: A look back and look ahead. *Perspectives in Health Information Management*, 19(1), 2022.

L. I. Kuncheva and C. J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine learning*, 51(2):181–207, 2003.

J. Leder-Luis. Can whistleblowers root out public expenditure fraud? evidence from medicare. *Forthcoming in The Review of Economics and Statistics*, 2023.

F. T. Liu, K. M. Ting, and Z.-H. Zhou. Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE, 2008.

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):56–67, 2020.

W. Luo and M. Gallagher. Unsupervised drg upcoding detection in healthcare databases. In *2010 IEEE International Conference on Data Mining Workshops*, pages 600–605. IEEE, 2010.

MedPAC. Hospital acute inpatient services payment system, 2023. URL `https://www.medpac.gov/wp-content/uploads/2022/10/MedPAC_Payment_Basics_23_hospital_FINAL_SEC.pdf`.

G. Nam, J. Yoon, Y. Lee, and J. Lee. Diversity matters when learning from ensembles. *Advances in Neural Information Processing Systems*, 34:8367–8377, 2021.

Noridian Healthcare Solutions. Unified program integrity contractor (upic), 2022. URL `https://med.noridianmedicare.com/web/jddme/cert-reviews/upic`.

T. Pevnỳ. Loda: Lightweight on-line detector of anomalies. *Machine Learning*, 102(2):275–304, 2016.

ResDac. Major diagnostic category (mdc) code, 2022. URL `https://resdac.org/cms-data/variables/major-diagnostic-category-mdc-code`.

M. A. Rosenberg, D. G. Fryback, and D. A. Katz. A statistical model to detect drg upcoding. *Health Services and Outcomes Research Methodology*, 1(3):233–252, 2000.

S. Sathe and C. C. Aggarwal. Subspace outlier detection in linear time with randomized hashing. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 459–468. IEEE, 2016.

M. Shi. Monitoring for waste: Evidence from medicare audits, 2022. URL `https://mshi311.github.io/website2/Shi_MedicareAudits_2022_09_15.pdf`.

E. Silverman and J. Skinner. Medicare upcoding and hospital ownership. *Journal of health economics*, 23(2):369–389, 2004.

N. C. Suresh, J. De Traversay, H. Gollamudi, A. K. Pathria, and M. K. Tyler. Detection of upcoding and code gaming fraud and abuse in prospective payment healthcare systems, Mar. 4 2014. US Patent 8,666,757.

U.S. Department of Health and Human Services. Annual report of the departments of health and human services and justice, 2022. URL https://oig.hhs.gov/publications/docs/hcfac/FY2021-hcfac.pdf.

U.S. Government Accountability Office. Payment integrity federal agencies' estimates of fy 2019 improper payments, 2020. URL https://www.gao.gov/assets/gao-20-344.pdf.

Appendices: For Online Publication

A Data Preprocessing

Our analysis of provider behavior uses data from each hospitalization and patient in the Medicare system. We consider patients hospitalized in 2017, and we use data from 2012 through 2016 to construct the patients' medical history.

A.1 Processing inpatient hospitalizations

We use 100% of samples of Fee-For-Service inpatient claims file from the Medicare data. Annual files contain beneficiary hospitalization details including provider, assigned DRG, assigned ICD codes, and payment reimbursement details including total payment amount, disproportionate payment, education payment, and outlier amount. The raw data is filtered to include claims where the total payment is greater than individual components. For example, if a claim has higher disproportionate payment compared to total payment amount, we exclude such a claim record from our data. These claims may indicate corrupted or noisy data recording. Next, to meet cell-size suppression requirement under our data agreement, we exclude providers along with their claim records, who served 10 or fewer beneficiaries in 2017. We then create lists of unique providers and beneficiaries from the filtered data, which we utilize for merging with other Medicare files.

A.2 Provider profile

First, we merge the filtered data with the master beneficiary summary files which contain beneficiary enrollment information including the beneficiary's address, demographics, and chronic conditions. Next, the data are merged with a DRG to MDC mapping.

We then create three types of provider representations. First, we collect the counts for each unique ICD code used by a given provider, creating a representation in terms of ICD codes used. This is a very high dimensional representation, where we apply our subspace based methods.

Next, for each provider, the counts of unique MDC codes are recorded. Since, each MDC typically corresponds to a part of the body, the MDC representation of providers gives a summary distribution in terms of the type of care they provide. Further, we collect counts of chronic conditions for each provider, which represents the distribution of patient population being served by a provider.

We also create the distribution over DRG codes for each provider by collecting the counts of unique DRG codes used by providers. This representation allows us to understand the spending pattern of a provider, since under the PPS system, the DRG code is directly tied to spending amount in each claim.

A.3 Beneficiary medical profile

In order to create a beneficiary's medical profile, we stitch through the patient's health care claims across different touchpoints in the Medicare system over the 5 years preceding the 2017 hospitalization (2012 – 2016). Specifically, for these years, we use 100% of samples of Fee-For-Service inpatient and outpatient claims, and 20% of samples of carrier files, which describe physician office visits. 20% is the largest available size of carrier files.

Given the volume of the datasets, we first filter the patient's visits across datasets based on the unique beneficiary list created from inpatient hospitalizations in year 2017. For each type of visit i.e. physician, outpatient, inpatient, we find unique diagnosis codes across five years. Next, for a given beneficiary, we collect the counts over the last five years for each of the unique diagnosis codes. We also include chronic conditions from the year 2016 and the patient's zip code from the master beneficiary summary file. Thus, a beneficiary is

represented in terms of assigned codes from past visits, chronic conditions and zip code.

B DOJ Corpus

We scrape and download press releases containing the word 'Medicare' from the central DOJ and the Offices of the United States Attorneys (USAO), which reflect local DOJ branches. The base URLs used in scraping for the DOJ and USAO are `https://www.justice.gov/news?keys=medicare` and `https://www.justice.gov/usao/pressreleases?keys=medicare`[3] respectively.

Next, we obtain the list of inpatient hospitals in the Medicare system from 'Medicare Inpatient Hospitals – by Geography and Service' dataset available from the Centers for Medicare and Medicaid Services at `https://data.cms.gov/provider-summary-by-type-of-service/medicare-inpatient-hospitals/medicare-inpatient-hospitals-by-geography-and-service`. This contains the information on providers including name, CCN (hospital ID), city, and state.

To find providers that are named in DOJ or USAO press releases, we first run a named entity recognizer[4] to obtain the names of all organizations from the press releases. We then run an exact name matching scan for each hospital in the list of Medicare inpatient providers in the recognized organizations from the press releases. Matched hospitals are then recorded as our ground truth. Next, we also run a partial name matching. We obtain tokens for each inpatient hospital in Medicare after dropping the word "hospital" in their name. Then we find organizations from scraped press releases that contain the tokens for Medicare hospitals. Since, we are matching tokens, multiple organizations match for a given Medicare hospital. We manually filter the multiple match and validate the match. The ground truth is augmented with our validated matches, which forms our DOJ corpus

[3]Webpages were accessed on accessed Mar 21, 2022.

[4]We used off-the-shelf entity recognizer Spacy available at `https://spacy.io/api/entityrecognizer`

for evaluation.

C Subspace detectors

(i) Subspace Outlier Degree (SOD) (Kriegel et al., 2009) locally examines each point (hospital) in the data. For each data point, it computes reference points through shared nearest neighbors. The subspace is then characterized by dimensions with low variance, lower than a provided threshold, within the identified reference set. It records the deviation of each data point from the hyperplane spanned by the mean of the identified subspace, where outliers have larger deviation.

(ii) Isolation Forest (IF) (Liu et al., 2008) builds a collection of randomized trees that approximate the density of data points in a random feature subspace characterized by paths in what are called "isolation trees". Each isolation tree is constructed by recursively partitioning data using a randomly chosen point in a randomly selected dimension, until the leaf of the tree contains a single data point. Shorter paths in a tree indicate sparse regions as fewer partitions lead to leaf nodes, and points belonging to each leaf at lower depth indicate outlierness in the subspace characterized by the tree path.

(iii) Robust Random Cut Forest (RRCF) (Guha et al., 2016), like IF, also constructs an ensemble of randomized trees by recursively partitioning the data. It computes the model complexity of each tree as the sum of the bits required to store the depths of each point in the tree. An outlier is defined as a point which increases the model complexity significantly when added to the tree.

(iv) Lightweight on-line detector (LODA) (Pevnỳ, 2016) constructs a collection of histograms on random 1-dimensional projections of the data. Each data point is then

associated with the negative log-likelihood based on each histogram, and data points are ranked based on their average likelihood across the 1-D histograms.

(v) RS Hash (RSHash) (Sathe and Aggarwal, 2016), like LODA, is also an ensemble of histograms; however, it constructs a collection of grid-based histograms in randomly chosen subspaces, and grid sizes vary based on varying sample sizes of data. Each data point is then scored by the number of sampled points sharing the same bin in the histogram. A sparsely populated bin is indicative of outlierness.

D Data Setup: Expenditure-Based Detection with Peer Analysis

Hospital representation.

We construct hospital profiles to capture the nature of services provided, the characteristics of patient population served, and encoding practices that drive spending for treatment.

Hospital profile – Type of services. We first examine a hospital's inpatient claims data to understand the type of services provided. Because the DRG codes assigned by hospitals may be manipulated to accomplish higher reimbursement, we must not represent hospitals by the exact DRGs they use; instead, we consider the hospitals's distribution into major diagnostic categories (MDC) (ResDac, 2022). Each MDC corresponds typically to one major body system (circulatory, digestive, etc), and can be associated with a set of medical specialties; each MDC contains a large set of potential DRGs. Therefore, characterizing hospitals by MDC allows us to consider hospitals that treat patients with similar types of medical needs, (e.g. digestive system issues), but without relying on the exact DRG codes assigned. For each hospital, we record the normalized count of each MDC code in the inpatient claims data in the current year.

Hospital profile – Patient population. We create another profile based on patient population characteristics served by a hospital. The underlying motivation for this profile is that two hospitals should be similar if they serve patients with similar medical conditions. To characterize the patient population at a broad level, we use the underlying chronic conditions of the patients. The chronic conditions flag whether a patient has received a previous set of services related to a chronic condition such as diabetes or ischemic heart disease. As a hospitals's representation, we record the normalized count of the chronic conditions of all the patients treated at the hospitals.

Hospital profile – Spending for care. The spending amount in each claim is directly tied to the assigned DRG code. To capture the DRG encoding practices of a hospital, we represent each hospital using the normalized counts of DRG codes from its inpatient claims. The DRG frequency representation allows us to compare and contrast the spending between a hospital and its peers that provide similar services or serve similar patients.

E Evaluation: Expenditure-Based Detection with Peer Analysis

Figure 6 shows the distribution of pairwise similarities between hospitals, and mark the similarity threshold at $\tau = 0.8$ which is used in our implementation for identifying peers. We exclude hospitals from our analysis that have less than five peers for the chosen threshold, as the estimation of excess spending could be noisy for these hospitals due to small peer group. Hospitals with large excess spending are ranked at the top and are identified as suspicious.

We use the DOJ corpus to evaluate our ranking of the providers based on excess spending. Figure 7 reports the PR and Lift curves for our peer analysis. The ranking is also compared to the two baselines, respectively ranking providers by average total claim amount

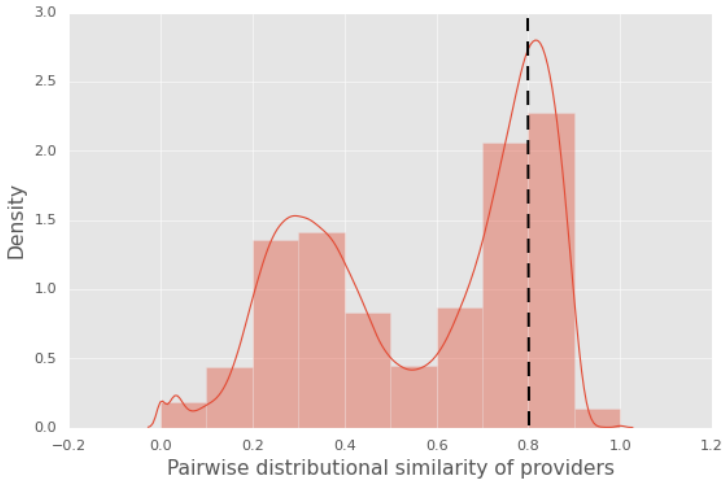


Figure 6: Distribution of pairwise similarities between hospital representations. A hospital and its peer hospital pair has similarity $>= 0.8$.

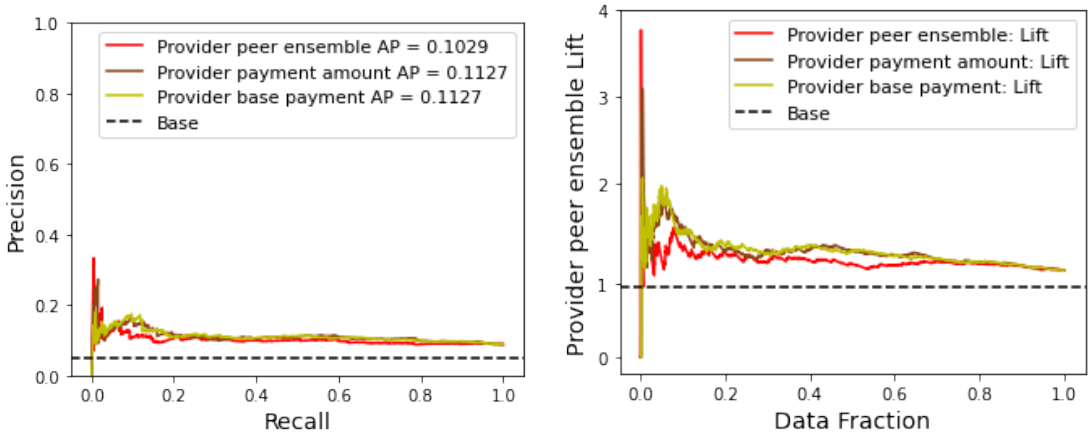
and average base payment amount. Although the peer-based ranking performance is comparable to these simple baselines, we remark that it is the lower bound on the performance. Furthermore, besides a mere ranking and unlike these simple baselines, our model can provide a nuanced explanation through DRG code frequency discrepancies, providing auditors with reasoning for potential factors driving the high spending. Finally, our model fundamentally identifies expensive hospitals as compared to their peers, which may be of interest to auditors interested in waste that may not rise to the level of fraud detected by the DOJ.

F Expenditure-Based Detection

F.1 Data Setup

Base payment amount.

For our analysis, we use the base payment amount computed from the Medicare inpatient claims. As explained in Section 2, the Medicare Prospective Payment System adjusts the claim payment amount to include expenses due to hospital variables such as patient mix,



(a) Precision-recall curve

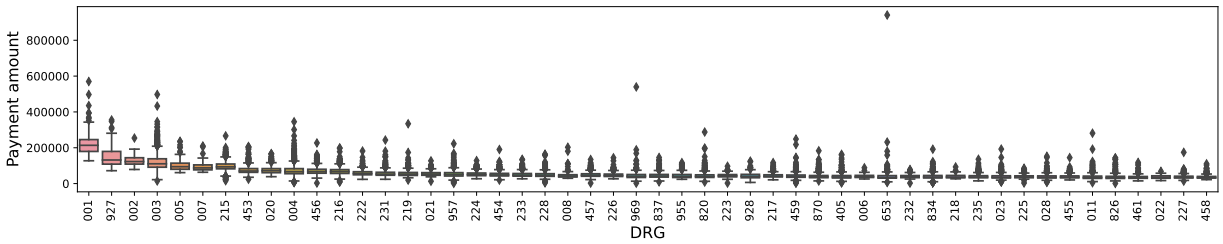
(b) Lift curve

Figure 7: We report the performance of ranking based on excess spending amount compared to the peers, where peers are identified via similarity based on MDC distributions and patient chronic conditions.

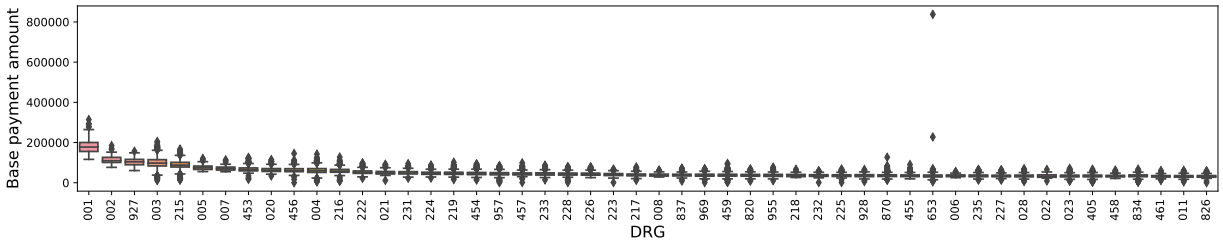
disproportionate share of low-income patients, outlier cases, and expenditure on education and research. These factors are generally external to the hospital's coding choice and should be excluded from analysis [5]. Therefore, to understand hospital behavior with respect to inpatient encoding, we rely on the base payment amount. The base payment amount is calculated by subtracting the reported adjustment amount from the total claim amount. While payments are also adjusted by hospital location through a geographically indexed wage, we do not control for wage index adjustments, because the geographical factor will be picked up when controlling for patient location in our regression.

Figure 8a shows the box plot of average total claim amount per hospital in the inpatient claims data (year 2017) for the top 50 DRGs, sorted by the mean of the box plot. Notice

[5]While previous work has shown that hospital outlier payments were subject to manipulation, those frauds ended in the early 2000s, about 10 years before the current study period. Moreover, those frauds did not happen at the patient or claim level, but rather involved hospitals manipulating other cost reporting documents



(a) Box plot of average *total claim amount* for DRG across hospitals from inpatient claims



(b) Box plot of average *base payment amount* for DRG across hospitals from inpatient claims

Figure 8: We plot the distribution of total claim and base payment amount across hospitals from inpatient claims in the year 2017. (a) Distribution of average total claim amount per hospital for the top 50 DRGs sorted on the mean of the box plot. There is large variation in the average claim amount for each DRG. (b) Box plot of the average base payment amount across hospitals for top 50 DRGs sorted on mean of the box plot. The variance across hospitals is lower for the base payment amount.

that there is large variation in the average claim amounts for each DRG. This variance across hospital is reduced when the box plot instead uses the average base payment amount as shown in Figure 8b. However, there remains some variance across hospitals even when considering the base payment amount.

Patient representation.

We represent each patient by their medical history and their covariates including location.

We consider all patients from 2017 who had an inpatient hospitalization claim and are at least 70 years old. Because Medicare is available for individuals aged 65 and older, we

include patients aged 70 years or above to ensure we observe a full 5-year history. We construct the medical history based on a patient's hospital visits in the previous five years (2012 - 2016). We filter and join patients data from physician visits, outpatient visits, and inpatient hospitalizations in the previous five years. Each patient visit, to a physician or inpatient facility, is assigned codes based on the ICD diagnosis and treatment codes. Thus, for a patient, we collect all the unique codes that were assigned in any of the visits along with their counts.

In addition to the treatment codes, we include the chronic conditions that require regular care, associated with each patient as reported in 2016, the year before the current year. We do not include 2017 chronic conditions as those may be outcomes of the code that the hospitalizations report. Including the 2016 chronic condition of a patient helps understand any comorbidities that may arise due to their medical history and ongoing chronic condition, accounting for the increase in treatment expense. Chronic conditions include diseases such as diabetes, breast cancer, Alzheimer's disease, and more. Our data provide a comprehensive view of the past treatments received by a patient, and reflects on their health. Further, to account for variation due to a patient's choice of hospital, as well as geographic differences in hospital reimbursement rates, we include the patient's location, represented by the first three digits of their zip code.

F.2 Detection Model

To estimate expected treatment expense for a patient, we employ a fixed-effects regression model with the outcome or target variable as the total base payment, and the features being the aforementioned patient representation (medical history and location).

We then include as regressors variables corresponding to the count of hospitalizations for that patient at each hospital. The coefficients of the hospital variables from this regression give the hospital fixed effects – in per hospitalization terms– that we use to rank hospitals.

Note that, because we are interested in capturing the hospital-level dependency of cost, we do not include treatment codes from the current year's hospitalization. The codes of the current year's hospitalization reflect the hospital's coding decision, which can be an element in its fraud or overbilling behavior. We address those in Section 5. Instead, the hospitals account for treatment expenses in the current year that are not reflected by the patient's medical profile; see Figure 1(a).

Regression model specification for expenditure.

Given (i) patient representation $\boldsymbol{X} \in \mathbb{R}^{N \times M}$ for N patients, each with a M-dimensional representation of historical medical profile based on the last five years (2012–2016), and (ii) the total base payment Y in year 2017; the specification for expected treatment expenditure prediction is as follows.

$$Y_i \;=\; \beta_0 \;+\; \boldsymbol{X}_i\,\boldsymbol{\beta} \;+\; \sum_j \alpha_j\, H_{j,i} \;+\; \epsilon_i \;\;, \tag{5}$$

where Y_i is the total base payment expense for a patient i in 2017; $\boldsymbol{X}_i$ is the patient representation for i, $\boldsymbol{\beta}$ depict regression coefficients associated with patient medical profiles and locations, $H_{j,i}$ is associated with an inpatient Medicare hospital j which contains total count of visits to j if patient i visited the hospital and 0 otherwise, and α_j's depict the hospital fixed effect regression coefficients.

Anomaly scoring.

In the expenditure-based regression, a coefficient α_j can be interpreted as the excess treatment cost due to hospital j that cannot be captured by patients' medical profile and location. As such, we can associate the magnitude and sign of this coefficient with the excess spending by a hospital, and designate it as its anomaly score.

F.3 Model Explanation

The regression model's hospital ranking in order of anomalousness is easily explainable through the coefficient values. Specifically, each α_j used for scoring and ranking has the direct interpretation as the excess expenditure on treatment for a patient when visiting hospital j. Therefore, the fixed effects model directly quantifies the excess dollar amount impact of a particular hospital, which can be used by an auditor or investigator when deciding which hospitals to investigate.

F.4 Evaluation

Figure 9 shows the estimated fixed effects, i.e. the α_j coefficients, for hospitals from our expected expenditure model. The hospitals with large fixed effects are ranked at the top and flagged as being of suspiciously expensive. In auditing, it is often the case that auditors have a limited budget (time and other resources) for processing red-flags and taking action. Thus, our method allows for targeting of audits towards the most suspicious hospitals, which corresponds to the highest unexplained spending.

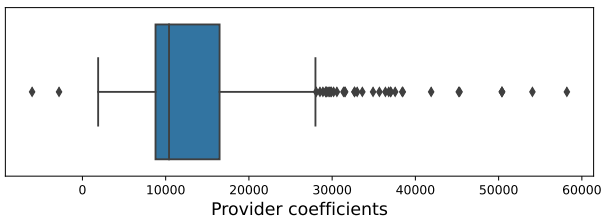
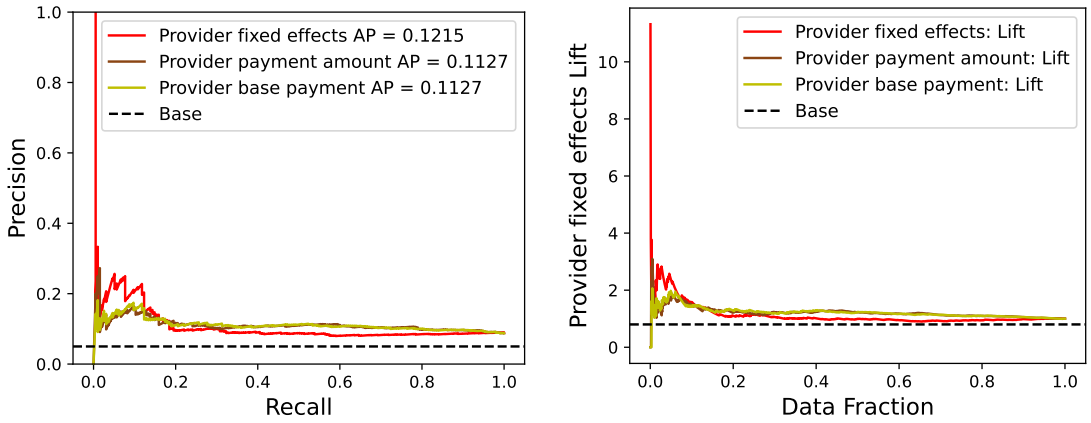


Figure 9: Distribution of the excess cost of treatment, that is, of α_j's in Eq. (5) per hospital j. The hospitals with large excess cost (coefficient) are ranked at the top for audit.

Figure 10 reports the PR and Lift curves for our fixed effects model, and compares its performance against two simple intuitive baselines. The baseline methods rank the providers based on average total claim amount and average base payment amount, respectively. Note that our fixed effects model is comparatively more effective at targeting

fraudulent hospitals, with relatively higher precision and lift at the top positions.



(a) Precision-Recall curve

(b) Lift curve

Figure 10: We report (a) Precision-Recall curve (AP: Average Precision denotes area-under-curve) and (b) Lift curve for provider ranking produced by fixed effects coefficients against two simple baselines: ranking of providers based on (1) average total claim amount and (2) average base payment amount. Dashed horizontal line 'Base' depicts the random ranking. Notice that top of the ranking is comparatively better as evidenced by higher precision and lift when recall and top data fraction are low. This is particularly helpful for auditors who would typically process only top ranked providers under limited budget.

Figure 11 reports the result of a two-sample test on the fixed effect coefficients as estimated by our model for providers in the DOJ corpus versus the rest of the providers. Notice that the DOJ providers typically have larger fixed effects as compared to others, and their distribution is significantly different as the test rejects the null that the two sets of coefficients are drawn from the same distribution, with $p < 0.001$. We remark that the reported performance is conservative and only the lower limit on our model's targeting ability, since many top ranked providers that are not part of DOJ ground truth may still have been involved in suspicious behavior. We report more qualitative results, and provide case studies through explanations into such flagged providers in Appendix G

, after accounting for the evidence from other models in our ensemble.

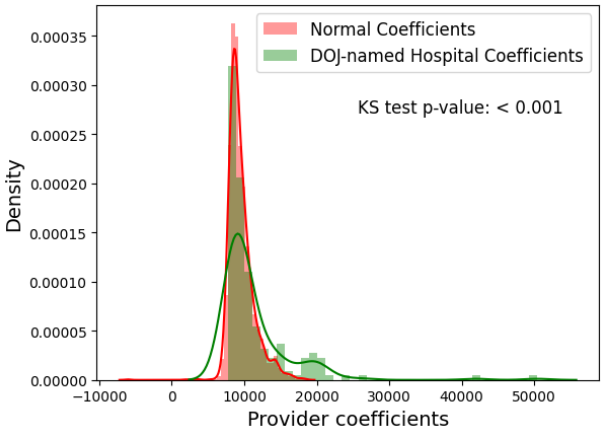


Figure 11: Comparison of fixed effect coefficients for providers facing anti-fraud lawsuits (known fraudulent entities or outliers) versus the rest of the providers (normal entities). A two-sample test rejects the null hypothesis, implying significantly different distributions statistically.

G Case Studies

In this section, we present an analysis of suspicious providers using our method, highlighting some of the salient aspects for the fraud detection task. We show how our method highlights parts of data from that contributed most to the ranking of a hospital as suspicious, which can assist in the process of auditing or deeper investigation.

We examine two top ranked hospitals from the aggregate ranking (1) the hospital at rank 1 that is also named in the DOJ corpus, and (2) the highest-ranked hospital which is not in our ground truth (at rank 5, as ranks 1–4 all are part of DOJ ground truth). In the following two case studies, we show how different models contribute evidence toward a better understanding of how each hospital stands out.

In Appendix H, we further provide model specific important ICD codes for the top 10 flagged hospitals. Notably, many of these ICD codes contain the word "other," indicating

they are less specific than related ICD codes and therefore more easily abused because of the ambiguity in their definitions.

Case 1: Flagged hospital named in DOJ corpus

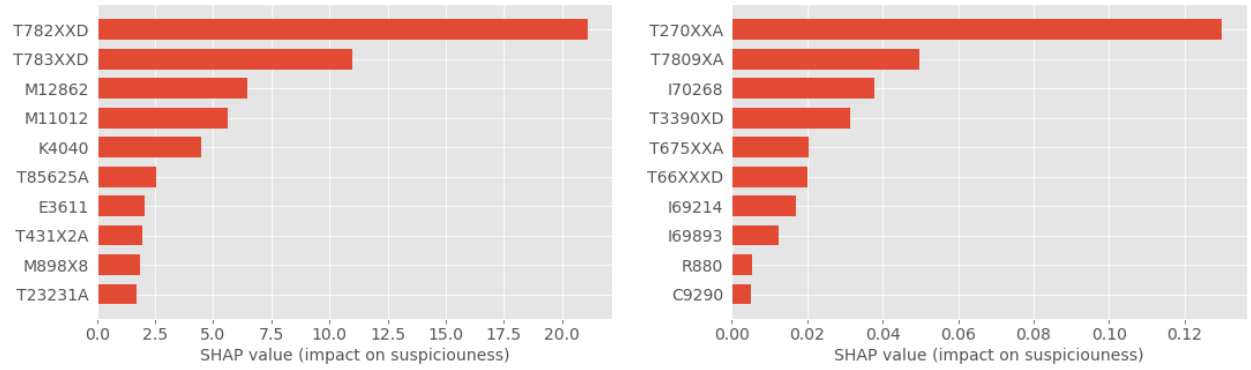
Our aggregate ranking finds the Cleveland Clinic as the most suspicious hospital under our metrics. Here we present evidence from our 3 Outlier Detection models, where this hospital is ranked at #1 by the subspace OD model, ranked at #17 by the peer-based model, and ranked at #27 by our regression-based model.

Notably, the Cleveland clinic settled with the DOJ in the years 2015 and 2021 for $1.74 million (DOJ Settlement, 2015) and $21 million (DOJ Settlement, 2021)[6] respectively. The evidence from our models do not directly match the reason for DOJ settlements; put differently, our exact explanations have not been validated externally by litigation. Moreover, our data do not provide evidence of fraud by the Cleveland Clinic, nor do they substantiate claims from lawsuits against the Clinic. The existence of previous lawsuits by the DOJ against the Clinic validate that this is a provider with past bad behavior, and our metric indicates that this hospital engaged in further anomalous behavior that can be detected by our algorithm and merits deeper investigation.

Figure 12a plots the most important ICD codes that contribute to the anomaly score of the hospital from the subspace OD model, based on SHAP values. The top ICD code "T782XXD" is described as "Anaphylactic shock, unspecified, subsequent encounter" which falls under the ancestor "T78" with the description: "Adverse effects, not elsewhere classified"[7]. As such, T78 appears to be a catch-all classification for adverse effects for injuries, poisoning, and other consequences of external causes for visit. Moreover, the code

[6]This 2021 enforcement was against Akron General Health System, which was acquired by the Cleveland Clinic foundation in 2015.

[7]ICD codes are available for lookup through ICD10Data. This code is available online at: `https://www.icd10data.com/ICD10CM/Codes/S00-T88/T66-T78/T78-`



(a) Top flagged hospital that is named in DOJ

(b) Top flagged hospital that is not in DOJ

Figure 12: ICD codes contributing to suspiciousness of top ranked hospitals based on SHAP values

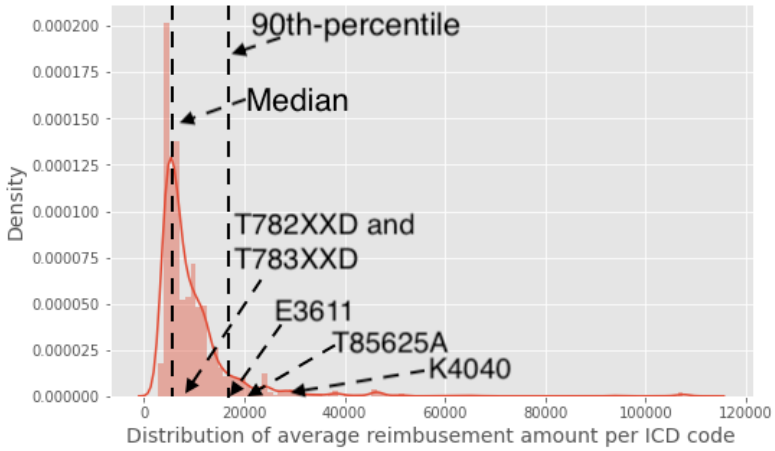


Figure 13: Distribution of ICD reimbursement amount obtained after mapping ICD code to most frequent DRG code in the inpatient claims data in year 2017. The median reimbursement amount is $6,650.88, and the 90-percentile reimbursement amount is $16,401.04.

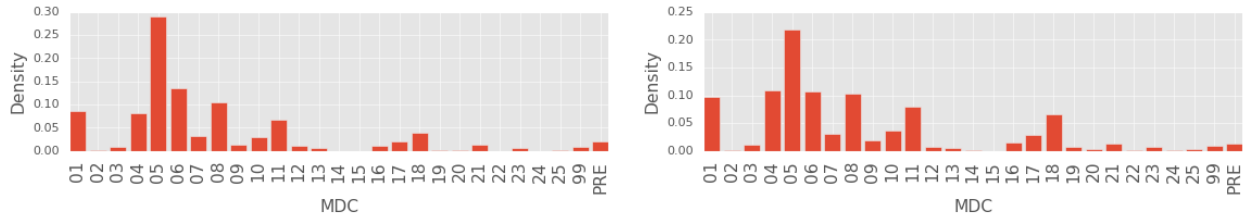
T782XXD is considered exempt from reporting whether the condition is present on admission (POA) to an inpatient facility. The next ICD code "T783XXD" is under the same ancestor, T78, and is also considered exempt from reporting if POA. Similarly, the description of code "M12862" allows non-specific reasons to be used for encoding as the given description is: "Other specific arthropathies, not elsewhere classified, left knee".

We next examine the reimbursement amounts related to these ICD codes, based on their mapping to the DRG they are most frequently associated with. The distribution of the amounts across all ICD codes is given in Figure 13. The codes T782XXD and T783XXD can be mapped to two DRG codes: 949 (Aftercare with cc/mcc) and 950 (Aftercare without cc/mcc).[8] The reimbursement amount for DRG code 949 is about 25% more compared to DRG code 950, where T782XXD is reported most frequently against DRG code 949. Further, within the ICD-10 hierarchy, codes T782XXD and T783XXD are the most expensive and get at least 50% more reimbursement than any other sibling or parent code. Notably, 6 out of top 10 ICD codes contributing to anomaly score (as shown in Figure 12) have reimbursement amounts that are more than 50th percentile among all ICD codes, while 3 of them associate with DRG codes with amount above the 90th percentile (see Figure 13). All these factors explain, through specific ICD codes, associated DRGs and dollar amounts, the reasoning behind why a flagged hospital stands out. This evidence provides starting points for further investigation.

In the peer-based model, the hospital is flagged through the peer relation of hospitals with respect to their MDC representation. Figure 14 shows the MDC distribution of the Cleveland Clinic and its nearest peer hospital. Notice that in terms of facilities and services provided as encoded by their MDC, the two hospitals are quite similar. We compare the DRG representation of the Cleveland Clinic to the summary DRG representation of all its peer hospitals over the top 50 DRG codes that are selected based on their contribution to excess spending (see Eq. 2 for excess spending estimate). As shown in Figure 15, Cleveland Clinic uses certain DRG codes more frequently than its peers as indicated by the summary distribution—starting with 219, 220, as well as 309, 310, 330. DRG codes 219 and 220 belong to "Cardiac Valve and Other Major Cardiothoracic Procedures" with reimbursement

[8]Here, 'cc' and 'mcc' stand for Complication or Comorbidity and Major Complication or Comorbidity, respectively.

amount in top 4 most expensive within MDC 05. DRG codes 309, 310 are described as "Cardiac Arrhythmia and Conduction Disorders", and DRG code 330 is described as "Major small and large bowel procedures with cc". Note that the description of codes 309, 310 and 330 is specific to a particular condition, while the description for 219–220 allows for ambiguity. Ambiguity may provide opportunities for miscoding to reach for higher reimbursement.



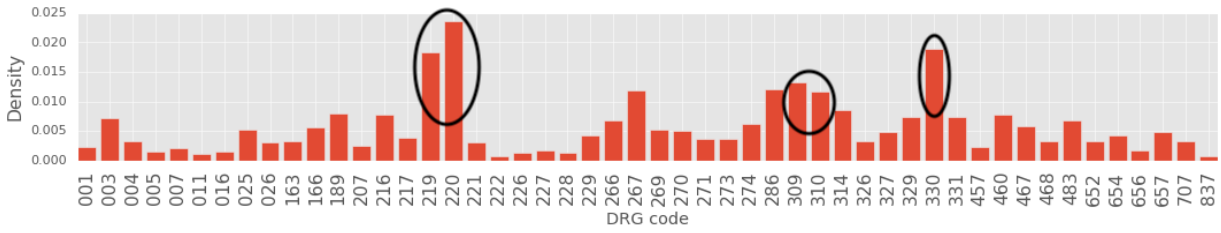
(a) Provider

(b) Nearest peer

Figure 14: Provider (named in DOJ) and its nearest peer represented in terms of MDC codes indicating provider facilities and services provided.

Our regression model estimates the excess expenditure on treatment for a patient when visiting the Cleveland Clinic to be $29,844.33, which is almost 3× the average expenditure (≈$10K) as shown in Figure 9. This does not, by itself, indicate that the Cleveland Clinic engaged in bad behavior, as this may reflect that it performs more specialized medical procedures, although our regression accounts for the patient's recent medical history.

One potential concern is that the hospital highlighted in this example, the Cleveland Clinic, as particular aberrant is a unique hospital that serves a particularly sick patient pool, and that therefore, the results are driven by selection of patients into different hospitals, as opposed to the effect of being treated at that hospital on expenditure. We argue this is not the case. Indeed, the two closest peer hospitals to the Cleveland Clinic are New York Presbyterian and Beth Israel Deaconess, both of which are similarly prestigious hospitals. Therefore, we expect that the results reflect actual divergent coding patterns by



(a) Provider (named in DOJ) DRG representation for MDC 05

(b) Summary DRG distribution of its peers for MDC 05

Figure 15: Comparing the DRG distribution of provider (named in DOJ) to the summary distribution created from its peer hospitals.

the most suspicious hospitals, rather than detecting hospitals that are engaged in specialty treatment.

In summary, all three outlier detection models point to evidence from different views of the claims data that makes the top ranked hospital stand out from others, both in terms of local and global analysis. These pieces of evidence explain the ranking by shedding light into certain coding practices that a hospital engages in, and may be utilized in further audit processes.

Case 2: Flagged hospital not in DOJ corpus

We now turn to a hospital which is flagged as suspicious by our metric but was never named by the Department of Justice in a press release about fraud.

In the aggregate ranking, AdventHealth Orlando hospital is ranked at #5 in order of suspiciousness. All 4 hospitals higher in the ranking were named in the DOJ corpus,

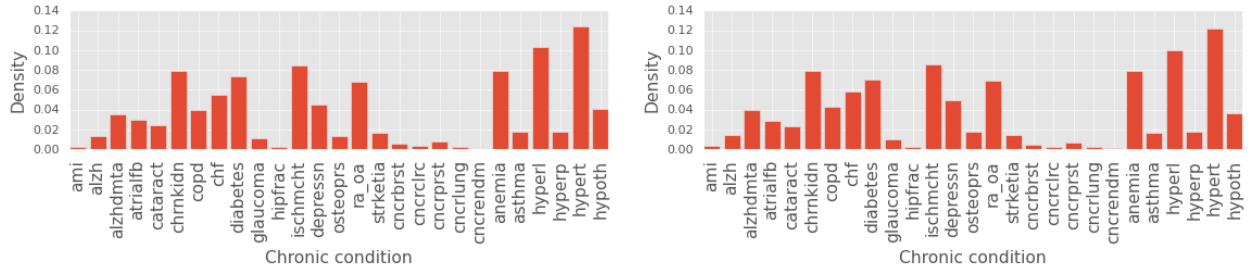
motivating this case study. This hospital is ranked at #5 by the ICD subspace model, and ranked at #35 by the peers-based model.

It is important to note that our model does not provide evidence of fraud, nor do we claim that AdventHealth Orlando has committed any fraud. Rather, our ranking of hospital suspiciousness can be used to guide further investigation and audits, and we use this case study to examine how our explainable model can help direct investigatory resources toward the exact claims that make a hospital different from its peers.

Figure 12b presents the bar plot of the top 10 ICD codes by importance for the hospital, based on SHAP values for the anomaly ranking from our subspace OD model. Note that 5 out of these top 10 ICD codes fall under ICD-10 chapter "S00-T88 Injury, poisoning and certain other consequences of external causes". The ICD code T270XXA is most frequently mapped to DRG code 205 which is described as "Other respiratory system diagnoses with mcc". The 3rd ranked ICD code "I70268" is described as "Atherosclerosis of native arteries of extremities with gangrene, other extremity". Based on the descriptions of these top ICD codes, a common thread appears to be that the codes leave room for ambiguity—due to the catch-all word 'other' in their descriptions. Further, 7 out of 10 ICD codes have reimbursement amount larger than the 50th percentile, and 4 out of 10 have reimbursements larger than 90th-percentile reimbursements across all ICD codes (recall Figure 13 for the ICD price distribution).

Next we present evidence from the peer-based OD model, though the hospital is not top ranked in this model. Figure 16 shows the hospital and its nearest peer hospital that serve similar patient populations, represented in terms of chronic conditions of the patients. We note the almost identical distributions of chronic conditions for the hospital and its nearest peer hospital. We compare the DRG distribution of the hospital to the summary DRG distribution of its peers.

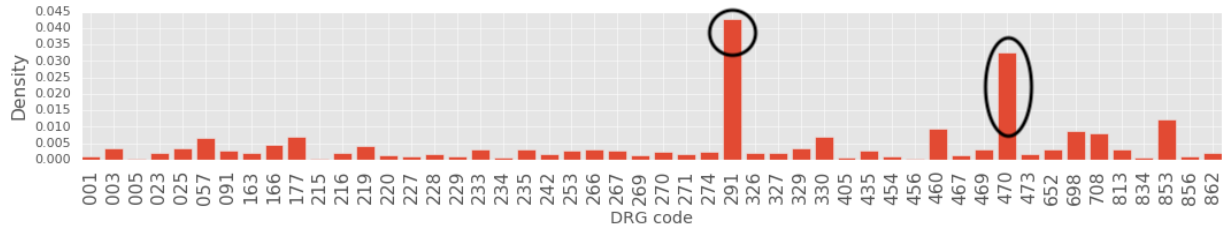
Figure 17 shows the distribution over the top 50 DRG codes, where the hospital's distri-



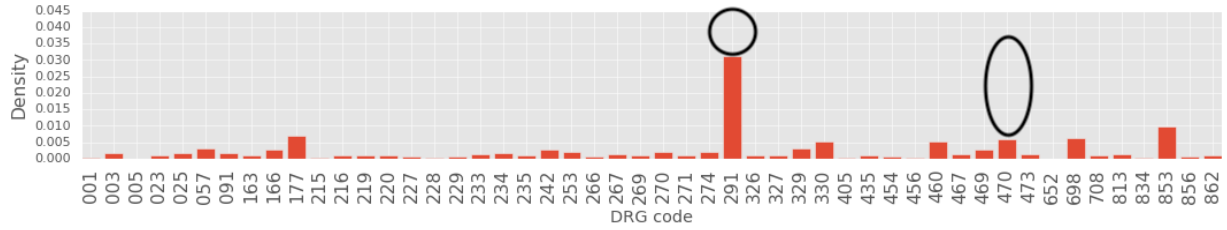
(a) Provider

(b) Nearest peer

Figure 16: Hospital (not in DOJ corpus) and its nearest peer represented in terms of patient population served



(a) Hospital (not named in DOJ) DRG representation.



(b) Summary DRG distribution of its peers

Figure 17: Comparing the DRG distribution of hospital (not named in DOJ) to the summary distribution created from its peer hospitals.

bution deviated from the summary distribution the most weighted by DRG reimbursement amount (see Eq. 2). We find that excess expenditure is almost entirely driven by two DRG codes, namely 291 (heart failure and shock with mcc) and 470 (major joint replacement or reattachment of lower extremity without mcc) with reimbursement costs larger than the 50th-percentile among DRG codes.

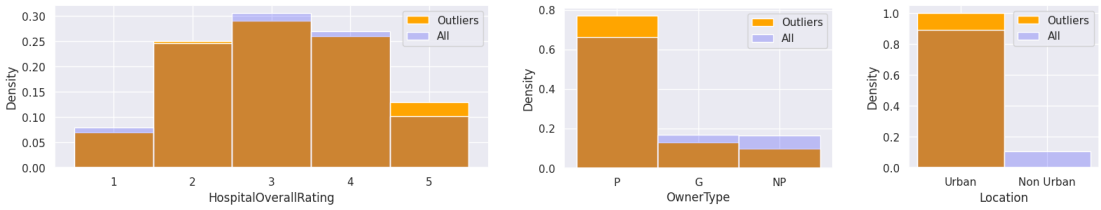
Similar to the earlier case, our models pinpoint specific ICD and DRG codes that can help jump-start further investigation, while highlighting dollar amount discrepancies that provide perspective with respect to monetary value.

Table 3: We show the top 2 ICD codes by importance for top 10 flagged hospitals. The ICD code description, and their most frequently mapped DRG codes are also included. The average payment amount for each DRG is reported. Notice that the flagging of hospitals is not driven by ICD codes for rare diseases, though some hospitals seem to include codes that have ambiguity in their description.

CCN	ICD1	ICD1 Desc	DRG1	DRG1 Desc	$ DRG1	ICD2	ICD2 Desc	DRG2	DRG2 Desc	$ DRG2	Patients	2017 Revenue ($)
360180	I8511	Secondary esophageal varices with bleeding	368	Major esophageal disorders w MCC	11282.86	C7651	Malignant neoplasm of right lower limb	829	Myeloprolif disord or poorly diff neopl w other O.R. proc w CCMCC	20735.67	11550	288M
340141	S5422XA	Injury of radial nerve at forearm level, left arm, initial encounter	074	Cranial & peripheral nerve disorders wo MCC	5563.80	S93115A	Dislocation of interphalangeal joint of left lesser toe(s), initial encounter	502	Soft tissue procedures wo CCMCC	7237.85	11855	208M
100128	O870	Superficial thrombophlebitis in the puerperium	776	Postpartum & post abortion diagnoses w/o O.R. procedure	5246.57	O10212	Pre-existing hypertensive chronic kidney disease complicating pregnancy, second trimester	781	Other antepartum diagnoses w medical complications	5287.21	6816	158M
220071	C227	Other specified carcinomas of liver	435	Malignancy of hepatobiliary system or pancreas w MCC	10720.54	M84462D	Pathological fracture, left tibia, subsequent encounter for fracture with routine healing	561	Aftercare, musculoskeletal system & connective tissue w/o CC/MCC	4621.62	12566	380M
100007	S72491B	Other fracture of lower end of right femur, initial encounter for open fracture type I or II initial encounter for open fracture NOS	481	Hip & femur procedures except major joint w CC	11431.83	M19172	Post-traumatic osteoarthritis, left ankle and foot	470	Major joint replacement or reattachment of lower extremity w/o MCC	11749.06	21951	401M

CCN	ICD1	ICD1 Desc	DRG1	DRG1 Desc	$ DRG1	ICD2	ICD2 Desc	DRG2	DRG2 Desc	$ DRG2	Patients	2017 Revenue ($)
140119	H66003	Acute suppurative otitis media without spontaneous rupture of ear drum, bilateral	153	Otitis media & URI w/o MCC	3872.24	M7611	Psoas tendinitis, right hip	558	Tendonitis, myositis & bursitis w/o MCC	4637.47	6400	177M
180056	D513	Other dietary vitamin B12 deficiency anemia	812	Red blood cell disorders w/o MCC	5167.92	S06351S	Traumatic hemorrhage of left cerebrum with loss of consciousness of 30 minutes or less, sequela	093	Other disorders of nervous system w/o CC/MCC	4013.56	889	10M
250004	D642	Secondary sideroblastic anemia due to drugs and toxins	812	Red blood cell disorders w/o MCC	5167.92	0SRR01A	Replacement of Right Hip Joint, Femoral Surface with Metal Synthetic Substitute, Uncemented, Open Approach	470	Major joint replacement or reattachment of lower extremity w/o MCC	11749.06	8799	122M
330024	I8511	Secondary esophageal varices with bleeding	368	Major esophageal disorders w MCC	11282.86	D4121	Neoplasm of uncertain behavior of right ureter	657	Kidney & ureter procedures for neoplasm w CC	12191.69	9560	325M
100006	0Y9C3ZZ	Drainage of Right Upper Leg, Percutaneous Approach	603	Cellulitis w/o MCC	4744.29	O870	Superficial thrombophlebitis in the puerperium	776	Postpartum & post abortion diagnoses w/o O.R. procedure	5246.57	8589	162M

I Characterizing Outlier Providers



(a) Hospital rating (scale 1 to 5) (b) Hospital ownership type: Private, Govt, Non-profit (c) Hospital location

Figure 18: Comparison of distributions over categorical covariates for Outlier hospitals and All hospitals

Figure 18 shows the normalized histograms for categorical covariates – hospital rating, ownership type, location type – for the hospitals. We compare the distributions for the top 5% of suspicious hospitals in aggregate outlier ranking with those over all the hospitals. The idea is that, assuming fraud is rare, an investigator with limited resources would examine only the top portion of the ranked hospitals.

We observe in Figure 18a that histograms for Hospital Overall Rating largely overlap, indicating that outlier hospitals and all the hospitals are sampled from a similar underlying distribution, i.e. hospital rating is *not* a strong predictor of outlier status. On the other hand, Figures 18b and 18c show that our top ranked fraudulent hospitals are more likely to be private for-profit urban hospitals, and less likely to be non-urban, government-owned or nonprofit hospitals. This observation agrees with the literature on for-profit care, which has found distortions from this ownership structure (Gupta et al., 2021).

J Aggregate Provider Ranking: ER

We present our analysis for Emergency Room (ER) visits only. We consider all ER patients from 2017, and construct their medical history using data from 2012 through 2016, following the same process as outlined in Section 3. We utilize the 100% sample of inpatient claims data from the year 2017 to filter claims marked as emergency admission (ER claims), and identify corresponding beneficiaries (ER patients). The ICD codes from the filtered ER claims, including diagnostics and procedure codes, are used to represent hospitals for our subspace model. For the peer analysis, hospital profiles are then constructed to capture service types and patient populations. MDC codes from ER claims reflect the service types provided by hospitals to ER patients. Similarly, patient population characteristics are based on chronic conditions of ER patients.

Further, for regression analysis, we construct patient profiles encompassing medical histories and location data for identified ER beneficiaries.

The ER claims data is a smaller subset, and consists of approximately 2/3 of the total claims that are used in our main analysis. We repeat our method entirely, aggregating the 8 detectors using rank choice voting. Even on this limited and specific sample, for hospitals ranked within the top 50, we obtain 2.5× lift as compared to random auditing when evaluating DOJ targeting.

K Subspace ranked results: Statistical significance

We used the $\sqrt{\epsilon}$ statistical test proposed by Chikina et al. (Chikina et al., 2017) to assess the significance of our ranked results. The test is based on the concept of an ϵ-outlier in a Markov chain (defined below), where we want to show that the presented state σ_0 (our ranked result) is unusual for states drawn from the stationary distribution π of the Markov chain.

Definition 1 (ϵ-outlier)**.** *$\alpha_0 \in \mathbb{R}$ is an ϵ-outlier among $\alpha_0, \alpha_1, \ldots, \alpha_k \in \mathbb{R}$, if there are, at most $\epsilon(k+1)$ indices i such that $\alpha_i \leq \alpha_0$.*

The test is defined as follows.

Definition 2 ($\sqrt{\epsilon}$ test (Chikina et al., 2017))**.** *Observe a trajectory $\sigma_0, \sigma_1, \ldots, \sigma_k$ from the state σ_0 for any fixed k. The event that $\omega(\sigma_0)$ is an ϵ-outlier among $\omega(\sigma_0), \omega(\sigma_1), \ldots, \omega(\sigma_k)$ is significant at $p = \sqrt{2\epsilon}$ under the null hypothesis that $\sigma_0 \sim \pi$.*

We set $\epsilon = 0.001$, and we ran the test on the Markov chain obtained from our experiment data. The test statistic was significant at $p < 0.05$, which indicates that we can reject the null hypothesis that the ranked results were not significantly different from a random ordering. This provides evidence that our ranked results are statistically significant. We next present the details of the test.

Data

We use the beneficiary inpatient claims data, and the hospital representation data from the inpatient claims from year 2017. The details of data setup is given in Section 5.

Constructing Markov chain

We define the Markov chain on the claims data as follows:

1. Given the current claims data, randomly select a tuple (c_j, H_j), where c_j is a claim from hospital H_j.

2. Randomly select a tuple from nearby hopsital (c_k, H_k).

3. Swap the membership of c_j and c_k to hospitals H_k, and H_j respectively, if it results in valid claims data

Valid claims data

To ensure the valid construction of Markov chain, we enforce certain conditions. These criteria are designed to facilitate the possibility of a beneficiary seeking treatment at a nearby hospital, where a the hospital may engage in different coding practice. The conditions are as follows:

1. Proximity Criterion: Hospital H_k and hospital H_j must be located within a 25-mile radius (based on zipcodes) of each other. This allows patients the option of choosing either hospital for potential treatment.

2. Profile Matching: Nearby hospitals H_k and hospital H_j must each have at least one patient whose medical history profile closely matches that of the beneficiary under consideration. The objective is to maintain the patient distribution within hospital.

3. DRG Code: The claims associated with these similar patients should have same assigned DRG code. Note that distinct ICD code assignments may still result in the same DRG code.

Runs of chain

We run the permutation for $k = 2^{30}$ steps to obtain states in the chain representing valid claims data. However, due to the scale of the data, we test our algorithm on the provider representation obtained from this valid claims data permutation every 10^8 swaps.

Label Function

The $\sqrt{\epsilon}$ test uses a label function $\omega(\cdot)$ to assign a real value to a state in the Markov chain. For our experiments, we use average top-n rank for the top 50 flagged hopsitals by our

algorithm as label function, defined as follows.

$$\omega(\sigma) = \frac{1}{n} \sum r_{H_j}^{\sigma}$$

where, $n = 50$, H_j are the hospitals flagged in our initial ranking, and $r_{H_j}^{\sigma}$ is the rank of H_j under permutation state σ.

L Aggregate Provider Ranking Excluding Medicare Advantage Patients

We present our analysis for visits excluding Medicare Part C (HMO) patients. These beneficiaries are covered by third party Medicare Advantage plans, and so we do not observe the complete medical history for these patients in the Medicare claims data. Therefore, as a robustness check, we repeat our analysis based on the claims that exclude these patients. We exclude any beneficiaries that are covered under HMO for any months between the years 2012 and 2017. This rules out 651 thousand beneficiaries from our target year of 2017, comprising 3.2 million hospitalizations.

Following the process outlined in Section 3, we construct the patient profiles from the filtered claims data. We utilize the 100% sample of inpatient claims data from the year 2017 after removing beneficiaries with HMO coverage. The ICD codes from the filtered claims, including diagnostics and procedure codes, are used to represent hospitals for our subspace model. For the peer analysis, hospital profiles are then constructed to capture service types and patient populations. Similarly, patient population characteristics are based on chronic conditions. Further, for regression analysis, we construct patient profiles encompassing medical histories and location data for identified non-HMO beneficiaries.

We repeat our method entirely, and obtain a 5-fold lift among hospitals ranked within the top 50 for detecting hospitals named by the DOJ.